The following publication Z. Liu, C. Lyu, Z. Wang, S. Wang, P. Liu and Q. Meng, "A Gaussian-Process-Based Data-Driven Traffic Flow Model and Its Application in Road Capacity Analysis," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 2, pp. 1544-1563, Feb. 2023 is available at https://doi.org/10.1109/TITS.2022.3223982.

A Gaussian-Process-Based Data-Driven Traffic Flow Model and Its Application in Road Capacity Analysis

Zhiyuan Liu*, Cheng Lyu, Zelin Wang, Shuaian Wang, Pan Liu, Qiang Meng

***Abstract*—To estimate the accurate fundamental relationship in traffic flow, this paper proposes a novel framework that extends classical fundamental diagram (FD) models to incorporate more dimensions of traffic state variables and allow for the impact of the supply-side factors of roads. The proposed framework is suitable for real-time traffic management, especially in urban areas, due to its reliance on minimal assumptions, its flexibility in adapting to various data sources, and its scalability to higher-dimensional data. The Gaussian process (GP) model is adopted as the base model for learning the optimal mapping from these input features to traffic volume. To enhance the GP model, an in-depth analysis of the properties of its kernel and likelihood function is provided. To cope with the hyperparameter optimisation of the GP, a modified Newton method for GP-based traffic flow model is also designed, which can jump over regions with small gradients. Experiments based on simulation data demonstrate the ability of the proposed framework to capture complex relationships between traffic state variables and supply-side factors, and show its value for estimating dynamic road capacity.**

***Index Terms*—traffic flow, fundamental diagram, Gaussian process, road capacity, hyperparameter optimisation**

I. Introduction

SINCE the pioneering work by Greenshields [1], numerous researchers have focused on interpreting traffic phenomena and improving traffic management practices using various models, including fundamental diagram (FD), three-phase traffic flow model, and higher-order traffic flow models. These models are valuable in revealing either the relationship between traffic variables in ideal scenarios or the statistical relationship between them in reality. They can be further enhanced for real-world applications in terms of assumption rationality, data availability, and scalability. First, classical models like the FD model and many kinematic-wave models [2], [3] work on the assumption of an equilibrium state in which all vehicles constantly run at the same speed and with the same spacing [4], [5]. This assumption is appropriate for near-stationary highway traffic [6], but may not accurately represent other real-life situations, especially complex urban traffic [7], [8]. As shown in Fig. 1, one typical issue is the wide scatter observed from two urban traffic detectors [9], and the high stochasticity of traffic makes it hard to correctly validate the FD model. Many successful efforts in traffic flow theory have been made to model traffic in a non-equilibrium state [10], [11].

Second, many of the traffic variables used in classical traffic flow models are difficult to obtain accurately from commonly available traffic sensors and hence are often approximated by other biased measures. The three basic variables in the FD, namely flow q, speed v and density ρ, should be defined in the same space-time domain such that the equation $q = \rho v$ holds. Regarding the definition in the space domain, the flow q in the equation should be the ratio of total speed of vehicles on a road section to the length of the road section rather than the number of vehicle passages over a unit time period [12]. Similarly, the speed v in the equation should be space mean speed rather than time mean speed. However, these variables (especially space mean speed) cannot be directly measured unless the trajectories of all vehicles are available, which is a challenge in a large city considering the cost of video cameras. Therefore, a common practice is to resort to surrogate quantities collected by fixed sensors, such as loop detectors and license plate recognition (LPR) devices, as an approximation, e.g., using the harmonic mean of instant vehicle speed over the cross-section of a road for space mean speed. These surrogates are merely estimations of real values, and the difference between them can be large [13], especially in the urban traffic environment. Traffic data unavailability becomes a barrier for high-order models. Their demand for fine-grained vehicle trajectories remains unresolved at present due to cost concerns.

Third, classical models are limited to the three fundamental traffic variables (flow, speed, and density), often described in a

This study is supported by the Distinguished Young Scholar Project (No. 71922007) and Key Project (No. 52131203) of the National Natural Science Foundation of China

Zhiyuan Liu is with the School of Transportation, Southeast University, China (e-mail: zhiyuanl@seu.edu.cn).

Cheng Lyu is with the School of Transportation, Southeast University, China (e-mail: cheng.lyu@seu.edu.cn).

Zelin Wang is with the School of Transportation, Southeast University, China (e-mail: wangzelin23@seu.edu.cn).

Shuaian Wang is with Department of Logistics & Maritime Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: hans.wang@polyu.edu.hk).

Pan Liu is with the School of Transportation, Southeast University, China (e-mail: liupan@seu.edu.cn).

Qiang Meng is with Department of Civil and Environmental Engineering, National University of Singapore, Singapore (e-mail: ceemq@nus.edu.sg).

(*Corresponding author. Email: zhiyuanl@seu.edu.cn)

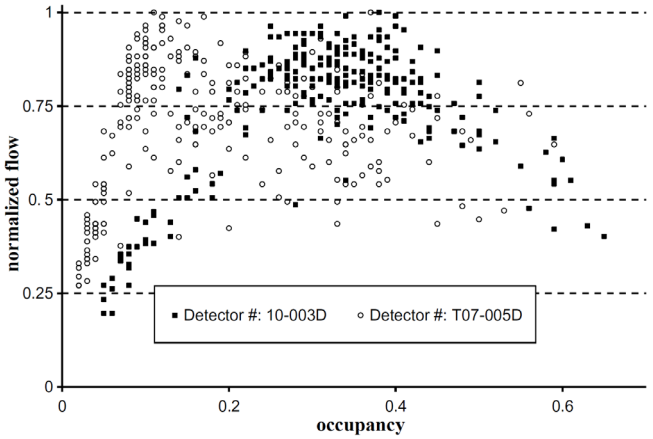


Fig. 1. The occupancy-flow scatterplot of two detectors [9]

bivariate functional form. In the era of big data, ubiquitous sensors for traffic data collection provide access to data of a much greater quantity and higher dimensions. However, the bivariate formulation restricts the utilisation of emerging data resources. It is worth noting that the traffic state is a joint outcome of diverse factors, e.g., speed, density, and vehicle type composition. Given the increasing amount of traffic data, there appears to be a need for a new model that can fully utilise these data to better portray traffic flow. In addition, only a few works on traffic flow have incorporated the influence of *road supply* on traffic. For instance, Jin and Zhang [14] modeled spatial inhomogeneity in road conditions. The external environment, including traffic signals and weather conditions, is often assumed to be temporally homogeneous during the data collection period (e.g., a week or month), but this assumption needs further consideration.

In this study, we, therefore, propose a data-driven model utilising big traffic data and machine learning technologies to overcome the three challenges of existing traffic flow models. The proposed model is applicable to cases where the assumption of an equilibrium state does not hold and leverages directly observable data. It is also scalable, allowing it to incorporate an extended set of flow-related variables and external factors influencing the supply of road resources.

A. Literature Review

The FD is a fundamental topic in the field of traffic flow theory and is the basis of many advanced macroscopic traffic flow models. It describes the relationship between flow, density and speed in the equilibrium state [4]. Research on FD models can be traced back to over 80 years ago when Greenshields conducted a photographic survey to study road capacity [1]. One typical formulation of the FD describes flow q as a function of density ρ, i.e., $q = f(\rho)$, and another popular formulation is speed v as a function of density ρ. These bivariate formulations are connected by the equation $q = \rho v$. FD is usually applied in assisting traffic control practices by determining the critical values of traffic state variables, e.g., capacity, critical density, and free-flow speed. In addition to road capacity estimation [1], [15], FD was also adopted as the theoretical foundation of ramp metering strategies like ALINEA [16]. Similarly, the design of the variable speed limit is also based on FD [17]. Moreover, FD is also useful as one of the basic inputs of LWR model and high-order traffic flow models for understanding the kinetic mechanism of traffic flow, where FD defines the relationship between volume, speed and density in the equilibrium state [2], [18]–[20]. Many studies have focused on exploring the functional form of the FD, which can be generally categorised as either single-regime or multi-regime. The single-regime FD models all traffic states, including both congested and uncongested states, using a single formula, such as Greenshields [1] and Newell [21]. Nevertheless, it is argued that the single-regime FD cannot adequately fit data in both congested and uncongested states [12], [22]. The multi-regime FD offers a solution to this problem by separately fitting data in different traffic states, e.g., Daganzo [23], Wu [24], and Li and Zhang [25].

Aside from the functional form of the FD, another line of research has focused on the estimation method for the parameters involved in these functional forms. The most straightforward and widely adopted parameter estimation method is least squares. However, this method has been criticised by Qu et al. [26] in that the uneven number of observations corresponding to traffic density may lead to biased estimations. Instead, it is more appropriate to adopt a weighted least squares method. In addition to generic parameter estimation methods, some researchers have proposed tailored methods to estimate each parameter of the FD based on its physical meaning. For instance, Dervisoglu et al. [27] designed a framework for estimating the triangular FD parameterised by free-flow speed, wave speed and capacity, in which the first two were estimated by performing linear regressions in two subsets of data, and capacity was estimated using historical maximum traffic volume. Similarly, Knoop and Daamen [22] attempted to separately calibrate each parameter of the FD proposed by Wu [24], including free-flow speed, wave speed, free-flow capacity, queue discharge rate and jam density. Both methods above rely on normal loop detector data. In contrast, Seo et al. [28] developed an algorithm to infer the FD based on the trajectories of probe vehicles, in which an FD for probe vehicles was estimated and then scaled using the jam density obtained from other data sources.

Emerging big data and machine learning technologies have attracted a lot of attention in a variety of fields. However, most research on traffic flow utilising these technologies has focused on traffic volume prediction. Following the seminal work by Ahmed and Cook [29], researchers have usually formulated traffic dynamics as a time series and extracted spatio-temporal patterns from historical data [30], [31]. Note that these predictive models focus on the relationship between volumes in different time instants rather than the relationship between different traffic flow characteristics like volume and density. In addition to traffic flow prediction, some researchers have managed to recreate network-wide traffic states from incomplete sensor data [32]. Moreover, machine learning models like neural networks have been introduced to microscopic traffic flow analysis to capture complex car following behaviours [33]. In recent decades, many researchers begin to exploit the capability of

Gaussian process (GP) in transportation modelling. Most of these GP-based studies aimed at traffic flow prediction. Focusing on network traffic volume forecasting, Sun et al. [34] made a comparison between various machine learning models, where GP shows its advantage in providing prediction results with uncertainty. Lin et al. [35] addressed the problem of traffic speed estimation on unobserved links using trajectory sensors on partial links and social media data, where GP is used to reconstruct the traffic speeds on all links. However, it is still an open question how to model macroscopic traffic flow using a data-driven approach, and a few researchers managed to address traffic state estimation with the help of GP in recent years. A very recent study by Yuan et al. [36] combined Gaussian process with classical macroscopic traffic flow models for traffic state estimation, but the third concern mentioned in the previous section remained unresolved. Li et al. [37] worked on traffic flow measurement based on floating car data, and GP was adopted to estimate traffic flow from traveling profiles. Sederlin et al. [38] developed their traffic state estimation model based on a Bayesian filtering framework, where GP was used to model the state-space transition of traffic states. However, these researches are not related to modelling the relationship between traffic state variables. Therefore, our research is intended to bridge this gap by developing a data-driven traffic flow model using GP regression.

B. Objectives and Contributions

Existing FD-based traffic flow models focus on the bivariate relationship among q, v and ρ. As mentioned above and by Hall [39], empirical observations rarely precisely accord with the fundamental equation $q = \rho v$. Thanks to the development of big data technology, a diversity of data from different sources can be collected. Therefore, in this study, higher dimensions of data are incorporated into the proposed framework, to provide an enhanced fitting of the empirical relationship of traffic variables. We seek to model the relationship in a data-driven way instead of beginning with a predefined hypothesis on the functional form of basic traffic state variables.

Another point often neglected in classical models is the influence of supply-side factors on the traffic state. We further attempt to capture their interdependency with traffic state variables by simultaneously modelling the internal factors (e.g., speed and density) and external factors (e.g., traffic signal) that determine the traffic state. Compared with classical FD models, the proposed model no longer requires recalibration for distinct road scenarios and can be generalised to unseen scenarios.

Because of the uncertainty of traffic flow, deterministic models are unsuitable for our goal. This research proposes the use of the Gaussian process (GP) model, a Bayesian machine learning model, to learn the complex nonlinear mapping among high dimensions of variables. To address the challenging issue of hyperparameter tuning of the GP, an improved search method called the modified Newton method for the GP-based traffic flow model is developed to reduce the risk of being stuck in a local optimum.

To sum up, this study aims at developing a new framework of traffic flow modelling that makes full use of big traffic data resources. Its contributions can be summarised as follows. First, the dimension of traffic state variables is extended compared with the classical bivariate formulation of the FD. Second, supply-side factors are also incorporated into the model, allowing our model to be applied to multiple road scenarios. Third, the use of the GP model for learning the complex relationship among input variables is proposed and a modified Newton method is designed to facilitate hyperparameter optimisation. It is worth noting that, as a data-driven black box method, the proposed framework is not a substitution for classical traffic flow models. Rather than interpreting the physics of traffic phenomena or predicting traffic state dynamics, it focuses on capturing the relationship among various traffic state variables and supply-side factors. As an example of its potential applications, we apply the proposed method to estimate the dynamic road capacity, which is a new concept that is valuable for urban road network analysis. Note that compared with the traditional FD-based approach, the proposed data-driven traffic flow model is more suitable for the analysis of complex and unstable traffic flows, especially for urban areas.

The remainder of this paper is organised as follows. We define the problem of learning fundamental traffic relationships and formulate the data-driven traffic flow model in Sections 2 and 3, respectively. A tailored machine learning model based on the GP and an improved searching method for hyperparameter optimisation are elucidated in Section 4. The proposed model is applied to estimate the dynamic road capacity in Section 5. Numerical experiments based on simulations are performed and analysed in Section 6. Finally, conclusions and future directions are provided in Section 7. A list of the main notations is provided in Appendix A for the sake of readability.

II. Problem Statement

Let us consider a set of road segments. To learn the relationship between traffic volume $\tilde{q} \in \mathbb{R}$ (whose definition will be discussed below) and an m-dimensional feature vector $\mathbf{x} \in \mathbb{R}^m$ derived from the data directly collected by sensors (for variables like vehicle speed) and surveys (for variables like road width), we aim to identify the mapping below using big traffic data,

$$f : \mathbf{x} \mapsto \tilde{q}. \tag{1}$$

By recording a set of vehicles V_T passing a specific road section during a period T, the volume $\tilde{q}$ can be defined as

$$\tilde{q} = \frac{|V_T|}{T} \tag{2}$$

where $|V_T|$ denotes the number of elements in set V_T. In the literature, $\tilde{q}$ is defined for a 'short roadway – long time' [2], [12], which is directly observable using a fixed traffic sensor. $\tilde{q}$ is different from another term, the flow q, which is used in some existing traffic flow models. q is defined for a 'short time – long roadway' [12], formulated as:

$$q = \frac{1}{L} \sum_{i \in V_L} v_L^{(i)} \tag{3}$$

where V_L denotes the set of vehicles on a road segment of length L at a specific time instant, and $v_L^{(i)}$ is the instant speed of the i-th vehicle in V_L. It can only be measured through overhead video cameras that continuously monitor a complete road segment. One critical problem in past research lies in the inconsistency between theoretical definitions and actually collected data. In this paper, $\tilde{q}$ is adopted for the definition of volume, instead of q, as it can be easily collected by most traffic sensors and is intuitive for administrators without hindering real traffic management.

The feature vector $\mathbf{x}$ in (1) can be decomposed into demand-side variables $\mathbf{x}_d \in \mathbb{R}^{m_d}$ and supply-side variables $\mathbf{x}_s \in \mathbb{R}^{m_s}$. In this paper, we mention traffic flow attributes like speed and density as demand-side variables to distinguish them from supply-side variables. These variables, along with volume, are all extracted from traffic flow itself, and their relationship can be captured by (1). Note that most classical traffic flow models only consider these demand-side variables, while neglecting the supply-side variables. The supply-side variables involved in (1) include factors related to roadway's geometric attributes and traffic control strategies like signal schemes. By incorporating these variables, variations in the relationship among demand-side factors and the differences among various roads can be simultaneously modelled.

A. Demand-Side Factors

Traffic volume is inherently correlated with demand-side variables such as speed and density. In real-time traffic operation and management, traffic flow data are obtained mostly from fixed sensors like loop detectors and LPR devices, which can provide highly accurate records of cross-sectional traffic variables. Both types of sensors are capable of reading instantaneous vehicle speeds. In addition, using loop detectors, the time occupancy can also be obtained, whilst with LPR devices, the vehicle type will be available in the dataset.

The two most important and widely adopted demand-side variables relating to volume $\tilde{q}$ are density ρ and speed v. Usually, individual observations of vehicles will be aggregated over a specific period, whereby statistics like average, variance, and median can be calculated to obtain density and speed. Regarding speed, the variable used in theoretical FD relationship should be space mean speed [40], that is, the arithmetic mean of all vehicle speeds in V_L, which cannot be directly measured by fixed sensors. Two common approximations to space mean speed are time mean speed $\tilde{v}_{\text{tms}}$ (the arithmetic mean of all vehicle speeds in V_T) and harmonic mean speed $\tilde{v}_{\text{hms}}$ (the harmonic mean of all vehicle speeds in V_T). Although it was proved that harmonic mean speed equals space mean speed in equilibrium state [40], they are not equivalent in reality. Likewise, density ρ in FD is also defined on V_L and cannot be directly measured by fixed sensors. As a result, it is often substituted by time occupancy o. The use of these approximations indicates that the measured traffic volume is not linearly correlated to the product of measured occupancy and speed. However, this issue can be resolved by introducing the data-driven model, which directly models the statistical relationship between the measured variables and the output without the equilibrium assumption.

B. Supply-Side Factors

The distributions of demand-side variables are conditional on supply-side factors. The geometric design of the road is a critical set of factors that determines the supply of road resources. Typical road attributes include road width rw, lane number lan, gradient g and median type mt. These variables place a constraint on the capacity of a road, and usually remain constant for a relatively long time.

In addition, traffic control and management methods like signal control and speed limit, denoted by sl, can potentially influence the traffic state. For instance, a traffic signal can impact road capacity, whilst the speed limit usually accords with the free-flow speed. Regarding traffic signals, these can be input into the model in the form of the green/cycle (g/C) ratio, denoted by gr. Due to the difference in traffic demand between peak hours and off-peak hours, gr may undergo several changes during the day, which will in turn affect traffic patterns.

These supply-side factors together constitute the vector $\mathbf{x}_s$. For example, $\mathbf{x}_s$ can include the elements of rw, lan, g, mt, sl, gr and can be further extended when more data are available.

III. Data-Driven Modelling of Traffic Flow

This section focuses on the formulation of the proposed framework for traffic flow modelling. The following two subsections discuss how the proposed framework extends the classical FD model from the perspective of the representation of the traffic state and the influence of supply-side factors.

A. Multi-Dimensional Representation of Traffic State

Recall the classical Greenshields FD, which can be formulated as

$$q = \rho v_f \left(1 - \rho / \rho_j\right) \tag{4}$$

where v_f denotes free-flow speed, and ρ_j denotes jam density. Similar to many FD models, it expresses flow as a function of ρ, as follows:

$$q = f_{FD}\left(\rho\right). \tag{5}$$

However, it is worth restating that the bivariate formulation of the FD is insufficient to account for the stochasticity of traffic flow, as pointed out by many studies [41], [42]. Recall the scatterplot in Fig. 1: wide scattering is often observed, which is not reflected by classical FD models. Wang et al. [43] dealt with this issue by proposing a stochastic modelling framework for the FD. Kerner [44] negated the rationality of the FD and promoted the three-phase partitioning of traffic flow. Another approach to this problem is a multi-class traffic flow model, which accounts for the heterogeneity of traffic flow in terms of the

speeds and lengths of different vehicle types [7], [45]. However, as mentioned earlier, higher-order traffic flow models and the three-phase flow model depend on fine-grained trajectory data, which are often unavailable in real applications. Moreover, the collected traffic data do not always conform to what we expect. For example, the speeds in many datasets are provided as time mean speed, but FD and kinematic-wave models require space mean speed.

Therefore, we propose to extend the FD model to include more factors, such that the uncertainty can be reduced and estimation accuracy can be improved. The proposed model can be formulated by

$$\tilde{q} = f_d\left(\mathbf{x}_d\right). \tag{6}$$

In this formulation, the features in $\mathbf{x}_d$ are not restricted to space mean speed and density, which are difficult to obtain, because their statistical relationship to volume can be directly captured by the model. For instance, $\mathbf{x}_d$ can be set as $[\tilde{v}_{\text{tms}}, o]$. In addition to density and speed in the classical FD model, more features can be exploited from data collected by various traffic sensors, such as small vehicle ratio $sr = \left|V_{T,s}\right| / \left|V_T\right|$, where $V_{T,s}$ denotes the set of small vehicles observed in period T. Then, $\mathbf{x}_d$ can be extended to $[\tilde{v}_{\text{tms}}, o, sr]$.

Another advantage of the data-driven formulation is its flexibility in adapting to various levels of data availability. When data resources are scarce, the proposed model can use basic features like occupancy and time mean speed. Accordingly, the model is more suitable for applications in real-time traffic management, compared with higher-order models restricted by the availability of fine-grained trajectories. Moreover, even if complete trajectory data become available in the future, the data-driven model is still advantageous as no strong assumption is made.

B. Incorporating Supply-Side Factors

Aside from extending the dimensionality of traffic state variables, we further incorporate the influence of supply-side factors in the model. The FD assumes that traffic and supply-side factors remain at a relatively stable state, which is rarely the case in reality, especially on urban roads. It should be stressed that traffic states are always constrained by the supply of road resources, and the distributions of traffic state variables are also conditional on this supply. Changes in supply-side factors will yield different traffic state patterns. A unified model incorporating both road supply and traffic state variables is able to elegantly address the above problem. Let the feature set of supply-side factors be $\mathbf{x}_s$; the new traffic model can be formulated as

$$\tilde{q} = f_{ds}\left(\mathbf{x}_d, \mathbf{x}_s\right). \tag{7}$$

Through this formulation, we can not only model the traffic flow on a road with time-varying road attributes but also on multiple roads simultaneously. Even for roads with constant physical attributes, a classical model without supply-side features has to be recalibrated for every road. Instead, the proposed formulation captures the latent relationship between various supply-side variables and traffic state variables. Thus, the traffic flow relationship can be deduced for unseen cases.

To identify the complex nonlinear relationship represented by f_{ds}, machine learning methods can be used instead of calibrating a simple predefined function as the FD model does. Although the proposed modelling framework is versatile enough to accommodate any machine learning model, it is expected that the selected model has an adequately large hypothesis space and is capable of dealing with data uncertainty. To achieve these expectations, GP regression is used as the base model for the proposed framework, because it has three merits.

- Unlike machine learning models that presume a class of function from input to output (e.g., linear regression), GP regression does not place explicit restrictions on functions it learns. It introduces a Bayesian view and places a prior distribution on each possible function, where higher possibilities will be given to those that better conform to the training data [46].
- The output of GP regression corresponding to each input vector is modelled as a random variable, and the learned function (i.e., the outputs corresponding to all input vectors) becomes a random process. The distribution of each random variable, on which the GP places a Gaussian prior such that the random process becomes easily tractable, can quantify the uncertainty of the output.
- A much smaller set of hyperparameters needs to be optimised for GP regression compared with complex models like Bayesian neural networks.

In the following section, we elaborate on how GP regression works and how it can be improved in the proposed modelling framework.

IV. Gaussian Process (GP) Regression

In this section, we give the definition and learning process of GP regression for traffic flow modelling. Part A and Part B discuss the basic procedures for GP regression and hyperparameter optimisation, respectively. For the proposed GP-based traffic flow model, the optimisation of the kernel hyperparameter is crucial but faces difficulty due to the nonconvexity of the optimisation model. Hence, in Part C, a modified Newton method for GP-based traffic flow model is proposed to address the issue.

A. Preliminaries of GP Regression

A GP is a random process, which gives the distribution of functions and can be adapted to solve regression problems [46]. Let $\mathrm{y} \in \mathbb{R}^{n\times 1}$ be a label vector of traffic volumes $\left[\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_n\right]^{\cdot}$, where n is the number of samples. The demand-side and supply-side features are represented by a feature matrix $\mathrm{x} \in \mathbb{R}^{n\times m}$, where m is the total number of demand-side and supply-side features. The aim of GP regression is to identify $f: \mathbb{R}^m \to \mathbb{R}$, that is, to obtain the relationship between traffic

volume and variables relating to supply-side variables and traffic state variables from the data. Note that the notation f is used to replace f_{ds} for simplicity.

Given a noise-free training dataset $\mathsf{D} = \{\mathtt{x}, \mathtt{y}\} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, where $y_i = f(\mathbf{x}_i)$, we denote the random variable corresponding to y_i by Y_i. A GP makes an a priori assumption that Y_i is normally distributed with mean $m(\mathbf{x}_i)$ and variance $\kappa(\mathbf{x}_i, \mathbf{x}_i)$, where $m(\cdot)$ and $\kappa(\cdot,\cdot)$ are a mean function and a kernel function, respectively.

For an arbitrary new input feature vector $\mathbf{x}_*$, the posterior distribution of Y_* follows a Gaussian distribution with mean m_* and variance Σ_*, formulated as

$$\begin{aligned} Y_* \mid \mathtt{y}, \mathtt{x}, \mathbf{x}_* &\sim \mathsf{N}\left(m_*, \Sigma_*\right) \\ m_* &= \mu_* + \boldsymbol{\kappa}_*^{\mathrm{T}} \mathbf{K}^{-1}(\mathtt{y} - \boldsymbol{\mu}) \\ \Sigma_* &= \kappa_{**} - \boldsymbol{\kappa}_*^{\mathrm{T}} \mathbf{K}^{-1} \boldsymbol{\kappa}_* \end{aligned} \tag{8}$$

where $\mu_* = m(\mathbf{x}_*)$ and $\kappa_{**} = \kappa(x_*, x_*)$. The covariances between $\mathbf{x}_*$ and every training sample are contained in vector $\boldsymbol{\kappa}_*$, the i-th element of which is $\kappa(\mathbf{x}_*, \mathbf{x}_i)$. The covariances between all training samples are stored in matrix $\mathbf{K}$, called the kernel matrix, whose element on the i-th row and j-th column is $\kappa(\mathbf{x}_i, \mathbf{x}_j)$. The posterior mean m_* here is a point estimation of Y_*, and the posterior variance measures the uncertainty of the estimation.

According to (8), it is obvious that the GP posterior is dependent on the choice of both prior mean function and prior kernel function, whilst the posterior variance is only affected by the prior kernel function. Usually, $m(\cdot)$ is directly set to be constantly zero, indicating that we assume at first that all inputs yield zero traffic volume. In this case, the posterior distribution of Y_* can be simplified to

$$Y_* \mid \mathtt{y}, \mathtt{x}, \mathbf{x}_* \sim \mathsf{N}\left(\boldsymbol{\kappa}_*^{\mathrm{T}} \mathbf{K}^{-1} \mathtt{y}, \kappa_{**} - \boldsymbol{\kappa}_*^{\mathrm{T}} \mathbf{K}^{-1} \boldsymbol{\kappa}_*\right). \tag{9}$$

Note that the kernel function plays an important role in the GP, which essentially acts as a similarity measurement between vectors in the input space. For traffic flow modelling, we expect similar input variables to correlate with similar traffic volumes. A common kernel to achieve this is the Gaussian kernel [47]. A standard Gaussian kernel is formulated as

$$\kappa_G(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2l^2}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2\right) \tag{10}$$

where l is the length scale of the kernel, which is the hyperparameter of the model.

The posterior mean of GP regression does not have a form with physical meaning like classical models. For example, analogous to an FD modelling the relationship between q and v, a zero-mean GP regression model with traffic volume $\tilde{q}$ as its label and speed $\tilde{v}_{\mathrm{hms}}$ as its only feature can be built. Recall that its posterior mean is $m_* = \boldsymbol{\kappa}_*^{\mathrm{T}} \mathbf{K}^{-1} \mathtt{y}$. We substitute the constant vector $\mathbf{K}^{-1}\mathtt{y}$ by $\mathbf{A} = [a_1, a_2, \ldots, a_n]^{\mathrm{T}}$. The posterior mean can then be written as

$$\begin{aligned} m_* &= \boldsymbol{\kappa}_*^{\mathrm{T}} \mathbf{A} = \sum_{i=1}^n a_i \kappa_G(\tilde{v}_{\mathrm{hms},i}, \tilde{v}_{\mathrm{hms},*}) \\ &= \sum_{i=1}^n a_i \exp\left[-\frac{1}{2l^2}(\tilde{v}_{\mathrm{hms},i} - \tilde{v}_{\mathrm{hms},*})^2\right]. \end{aligned} \tag{11}$$

The Gaussian kernel above writes the covariance between $\mathbf{x}_i$ and $\mathbf{x}_j$ as a function of $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2$. As a result, when more dimensions of features are incorporated, the covariance will still change in the same scale with respect to all features. In other words, the kernel is isotropic. To allow the model to adapt to various dimensions of features, the standard Gaussian kernel can be modified as,

$$\kappa_{G,\mathrm{ani}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2}\left(\mathbf{x}_i - \mathbf{x}_j\right)^{\mathrm{T}} \boldsymbol{\Lambda}\left(\mathbf{x}_i - \mathbf{x}_j\right)\right), \tag{12}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}(l_1^{-2}, l_2^{-2}, \ldots, l_m^{-2}), \tag{13}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix defining unique length scales for each dimension of input features (m denotes the dimension of features), enabling GP regression to automatically determine the importance of each feature and reduce the impact of irrelevant features [48]. A large length scale in a specific feature dimension indicates that this feature makes little contribution to the prediction, whereas a small length scale indicates an important feature in prediction.

B. Hyperparameter Optimisation of GP Regression

A zero-mean GP can be completely determined by its covariance matrix, i.e., the kernel matrix. Therefore, training a GP regression model is essentially a search for the optimal hyperparameters of the kernel. Here, we denote the vector of all hyperparameters of a kernel by $\boldsymbol{\theta}$. For example, the hyperparameter vector of an anisotropic Gaussian kernel described in (12) and (13) is $\boldsymbol{\theta} = [l_1, l_2, \ldots, l_m]$. The optimal hyperparameters $\boldsymbol{\theta}^*$ are obtained through maximising the log marginal likelihood (LML) of the GP regression, $\log p(\mathtt{y} \mid \mathtt{x}; \boldsymbol{\theta})$, which means that we would like to maximise the probability of observing $\mathtt{y}$ given the input data $\mathtt{x}$ and the kernel matrix $\mathbf{K}$ parameterised by $\boldsymbol{\theta}$. As a GP for a finite set of samples follows a multivariate Gaussian distribution, we have

$$\mathbf{Y} \mid \mathbf{x}, \boldsymbol{\theta} \sim \mathsf{N}\left(\mathbf{0}, \mathbf{K}\right) \tag{14}$$

where $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]^{\mathrm{T}}$, $\mathbf{K}$ is a covariance matrix parameterised by $\boldsymbol{\theta}$. Then, the LML of this GP regression can be formulated as

$$\begin{aligned} \log p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) &= \log\left[\frac{1}{\sqrt{(2\pi)^n |\mathbf{K}|}} \exp\left(-\frac{1}{2}\mathbf{y}^{\mathrm{T}} \mathbf{K}^{-1} \mathbf{y}\right)\right] \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\mathbf{y}^{\mathrm{T}} \mathbf{K}^{-1} \mathbf{y}. \end{aligned} \tag{15}$$

According to (15), LML comprises three terms, the first of which is a constant. Therefore, maximising LML is equivalent to minimising $L(\boldsymbol{\theta})$:

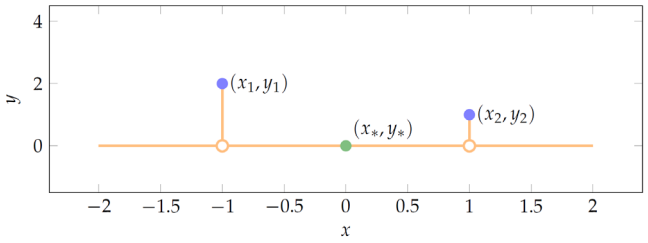


Fig. 2. Posterior mean of a GP regression with a small length scale (0.001) of Gaussian kernel

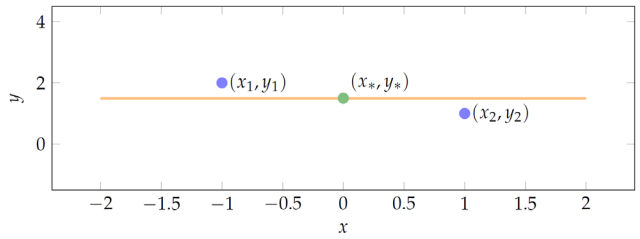


Fig. 3. Posterior mean of a GP regression with a large length scale ($_{10^0}$) of Gaussian kernel

$$L(\boldsymbol{\theta}) = \log|\mathbf{K}| + \mathbf{y}^{\acute{u}}\mathbf{K}^{-1}\mathbf{y}\,. \tag{16}$$

Among the two terms of $L(\boldsymbol{\theta})$, the former, i.e., $\log|\mathbf{K}|$, works as a regulariser corresponding to the model complexity for GP regression, whereas the latter, i.e., $\mathbf{y}^{\acute{u}}\mathbf{K}^{-1}\mathbf{y}$, reveals the data fitting quality.

Two extreme cases of the GP kernel matrix $\mathbf{K}$ are the identity matrix $\mathbf{I}$ and the matrix of ones $\mathbf{J}$, where the former leads to an extremely flexible GP model that can perfectly fit any given data. For illustration, consider two arbitrary samples as shown in Fig. 2. Clearly, the posterior mean of GP regression drastically increases near the two samples, but remains approximately zero for input x_* elsewhere. Whatever $\mathbf{y}$ is given, the GP will severely overfit the data. The other extreme case is associated with severe underfitting. In this case, the posterior mean of GP regression near sample points will be almost constant, as demonstrated in Fig. 3. For most input x_*, the estimated y_* is constant around 1.5.

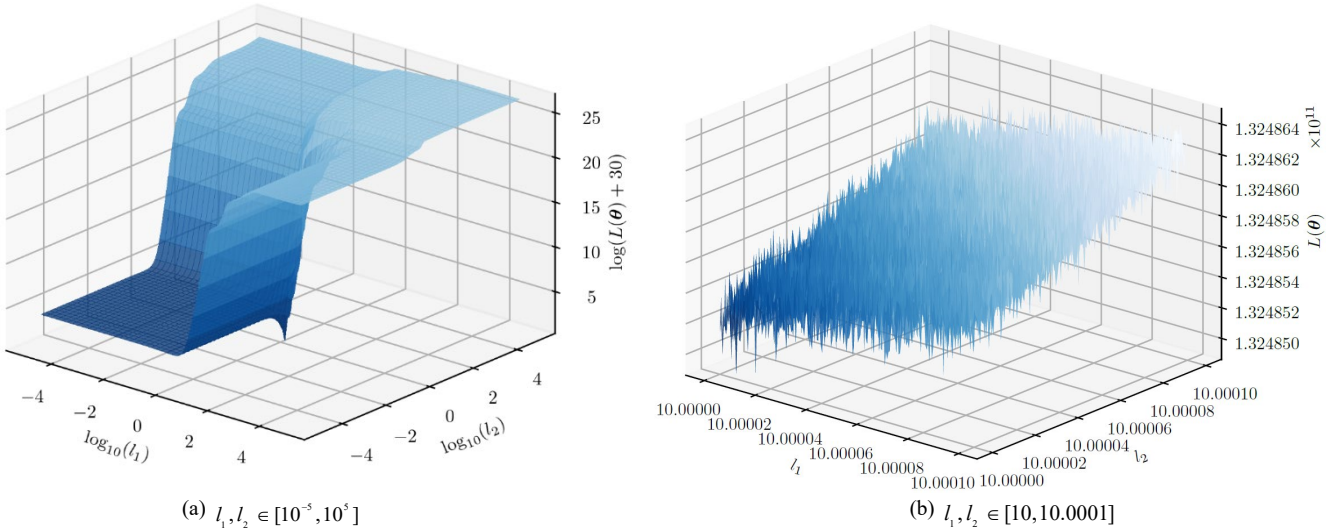
C. A Modified Newton Method for Hyperparameter Optimisation

An appropriate set of hyperparameters, i.e., the length scale of the Gaussian kernel, should be a careful trade-off between $\log|\mathbf{K}|$ and $\mathrm{y}^{\acute{u}}\mathbf{K}^{-1}\mathrm{y}$ to avoid overfitting and underfitting. To find the optimal hyperparameters, first find the derivative of $L(\boldsymbol{\theta})$,

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}}\left(\log|\mathbf{K}| + \mathbf{y}^{\acute{u}}\mathbf{K}^{-1}\mathbf{y}\right) \\ &= \mathrm{tr}\left(\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}}\right) - \mathbf{y}^{\acute{u}}\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}}\mathbf{K}^{-1}\mathbf{y}.\end{aligned} \tag{17}$$

Unfortunately, the optimisation objective, i.e., $L(\boldsymbol{\theta})$, is often nonconvex. For the sake of presentation, we take a two-dimensional sinusoidal function, formulated as $f_{toy} = \sin x_1 + \sin 3x_2$, as an illustrative example. We randomly sample 80 points between -5 and 5 to examine the $L(\boldsymbol{\theta})$ surface, $\boldsymbol{\theta} = (l_1, l_2)$, where l_1 and l_2 are the length scales corresponding to the two dimensions. The $L(\boldsymbol{\theta})$ with respect to length scales is plotted in Fig. 4 (a), where the z-axis is shifted and log transformed for the sake of demonstration. Fig. 4 (b) shows an enlarged view of part of Fig. 4 (a), where $l_1, l_2 \in [10, 10.0001]$.

In existing machine learning platforms (e.g., Scikit-learn [49]), gradient-based algorithms, like the Newton method, are commonly used for the $L(\boldsymbol{\theta})$ maximisation of GP regression. Considering the nonconvexity of $L(\boldsymbol{\theta})$, the multi-start Newton method is often adopted in the hope that a better initial point can be selected. However, we find that the gradient soon becomes zero and the optimiser is stuck in a local optimum, causing the overfitting of the GP regression model, as depicted in Fig. 5. In comparison, the model should reasonably fit the data



(a) $l_1, l_2 \in [10^{-5}, 10^{5}]$ (b) $l_1, l_2 \in [10, 10.0001]$

Fig. 4. $L(\boldsymbol{\theta})$ surface of the toy example ($f_{toy} = \sin x_1 + \sin 3x_2$)

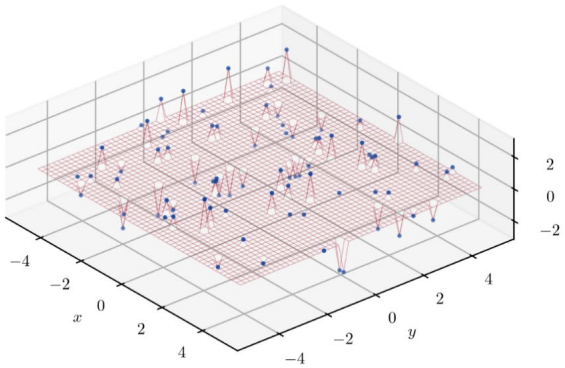


Fig. 5. Posterior mean surface of an overfitted GP with $L(\boldsymbol{\theta}) \approx 67.00$

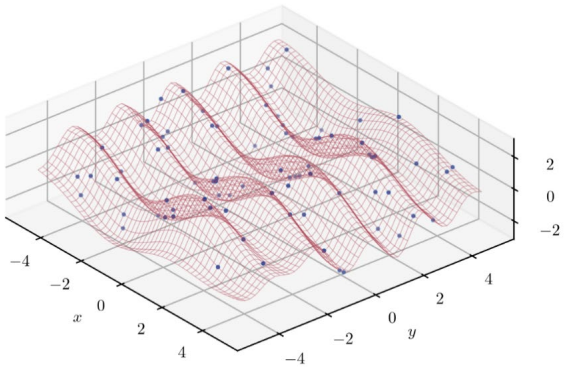


Fig. 6. Posterior mean surface of a reasonably fitted GP with $L(\boldsymbol{\theta}) \approx -144.47$

if the global optimum of $L(\boldsymbol{\theta})$ is reached, as depicted in Fig. 6. The reason for the failure of the common multi-start Newton method can be attributed to the characteristics of the $L(\boldsymbol{\theta})$ surface. For many hyperparameter sets, the $L(\boldsymbol{\theta})$ surface often appears to be a plateau where the gradient is extremely small, especially when all length scales of the anisotropic Gaussian kernel are small.

Proposition 1. For a GP with an anisotropic Gaussian kernel parameterised by a length scale vector $\boldsymbol{\theta} = [l_1, l_2, \ldots, l_m]$, where $l_i > 0$, $i = 1, 2, \ldots, m$, we have

$$\lim_{\boldsymbol{\theta} \to \mathbf{0}^+} \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{18}$$

The proof of Proposition 1 can be found in Appendix B.

Aside from the plateau of the $L(\boldsymbol{\theta})$ surface when $\boldsymbol{\theta}$ is small, the surface can be numerically unstable when $\boldsymbol{\theta}$ becomes larger. As demonstrated in Fig. 4 (b), we magnify the surface to the range of $l_1, l_2 \in [10, 10.0001]$. It is apparent that the surface is rough, which can be difficult for gradient-based algorithms to optimise.

Therefore, compared with directly performing a multi-start search, it is more reasonable to truncate the hyperparameter space with near-zero or unstable gradients in the outer rim. To achieve this goal, we propose a modified Newton method for the hyperparameter optimisation of GP regression. The algorithm comprises two stages: the first searches for the bound of the region with near-zero gradients, and the second is the ordinary multi-start Newton algorithm. The bound search begins

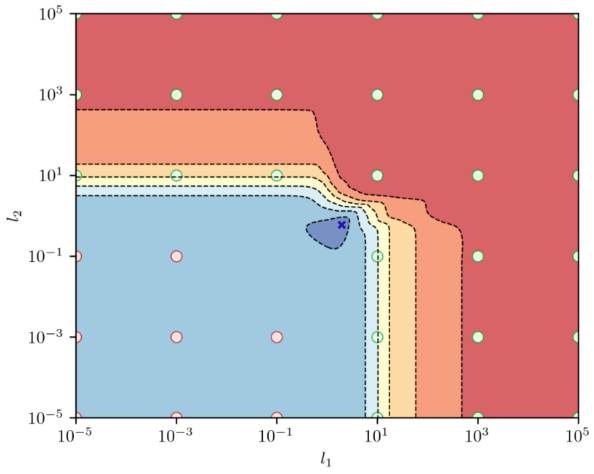


Fig. 7. Bound search of modified Newton method for GP-based traffic flow model

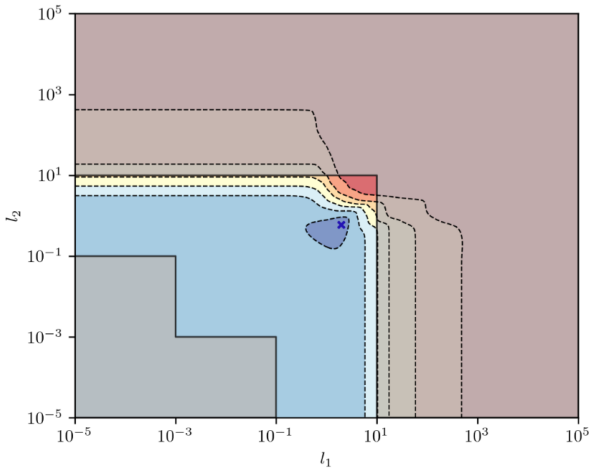


Fig. 8. Truncated space for hyperparameter search

from the lower and upper bounds of the length scales, respectively, to truncate the hyperparameter space. For example, a typical two-dimensional contour of $L(l_1, l_2)$ is illustrated in Fig. 7, where a darker blue colour indicates a higher value. The optimal hyperparameters are located in the centre of the Figure. It is highly possible that the algorithm will get stuck in the plateau when a starting point is randomly picked. Before searching, the hyperparameter space has to be discretised. For each dimension of the input feature, k values are selected as probes $\mathbf{b}$, which are equally spaced on a logarithmic scale within the predetermined lower and upper bounds.

We first search for the new lower bound of the length scales. The probes should be sorted in ascending order, indicating that the algorithm starts from the lower bound. As presented in Algorithm 1, Q is initialised in the beginning so as to store the truncated lower bounds. In addition, the value of $L(\boldsymbol{\theta})$ when all length scales are set to their lower bounds is first computed as L_0. Then, we iterate through all dimensions of the features and compute the value of $L(\boldsymbol{\theta})$ corresponding to each combination of length scales, i.e., $L(p_1, p_2, \ldots, p_m)$. If its difference from L_0 is greater than the threshold Δl, the combination of length

scales will be labelled as a bound point and added to Q.

The new upper bounds of the length scales can be searched in a similar way. However, note that the probes $\mathbf{b}$ should be sorted in descending order so that the search can start from the upper bound. Also, the bound condition in Line 6 of Algorithm 1 can be changed to $L(p_1, p_2, \ldots, p_m) - L_0 < \Delta l$ considering that $L(\boldsymbol{\theta})$ at the optimum should be smaller than L_0.

New lower and upper bounds can be used in substitution for the original ones when sampling the initial points for the multi-start Newton method, to narrow down the search region. For example, as illustrated in Fig. 8, the initial points should not be selected within the grey region.

V. Application in Dynamic Road Capacity Estimation

A. Dynamic Road Capacity Estimation

The proposed data-driven model can be adopted in many suitable applications, e.g., optimisation of road design. In this paper, for the sake of presentation, we take the road capacity analysis as an illustrative example to present the application of the proposed data-driven model. Meanwhile, a new concept termed *dynamic road capacity* is proposed and analysed, as a representative case to show the merits of the new model. The road capacity is extended from a stable value to a variational term, mainly because the dynamic change of supply-side factors, rather than the demand-side factors.

The notion of road capacity in this study refers to the maximum traffic throughput that a road allows to pass, similar to the notion of practical capacity in engineering practices. Road capacity is an essential ingredient in analysing and optimizing traffic plannings, designs, and management schemes [50]. Thus, there has been a long-standing and growing interest in estimating capacity. *Highway Capacity Manual* (HCM) provides some guidelines on the estimation of link capacity in engineering practices, whereas, the crude default values of various external factors limit its generalizability [51]. Besides following the guidelines in HCM, one of the most widely used capacity estimation methods is the FD-based method [52]-[55]. The FD-based methods construct traffic flow relationships, which are derived based on field data, to estimate road capacity [53], [56]. Many studies estimate capacity in freeways considering physical and environmental conditions [57]-[59]. In recent years, capacity has been extensively studied in urban networks, in which the effects of traffic facilities such as pedestrian crossings and signalized intersections are considered [60]-[64]. Existing analysis of road capacity generally revolves around a static definition of capacity oriented to road design. The capacity of a road is assumed to be constant over a relatively long period. Nonetheless, supply-side factors like traffic signals that influence capacity usually change over time, and correspondingly, the road capacity also varies over time. This is unfortunately not considered in the existing notion of saturation rate. Aside from dynamic signal controlling schemes, adverse weather can also restrict a road from reaching the same level of capacity as on sunny days. When the FD is adopted for capacity estimation, all historical traffic state observations are placed in the same diagram, making it impossible to tell the capacity variations. Therefore, a dynamic definition and a new approach to road capacity estimation are required. We believe that this definition can be valuable in engineering practices because it provides traffic engineers with a direct value of the maximum number of vehicles that can pass the road in prevailing cases. Based on this value, the network capacity can be further obtained, which is helpful in regional traffic control.

Based on the proposed data-driven traffic flow model, the road capacity in scenarios with different supply-side features can be easily obtained. Given the supply-side features $\mathbf{x}_s$ of a road, its capacity $c(\mathbf{x}_s)$ can be obtained by maximising the volume of (7),

$$c(\mathbf{x}_s) = \max_{\mathbf{x}_d} f_{ds}(\mathbf{x}_d, \mathbf{x}_s). \tag{19}$$

As the posterior of a trained GP regression model can be expressed in a closed form, the optimisation problem in (19) is easily tractable. The continuity of the GP naturally suggests the use of gradient-based algorithms for optimisation.

It is worth noting that the lack of ability of causal inference has long been regarded as a shortage of machine learning models, but it does not mean that these models cannot generalise to other cases. For various applications of machine learning models nowadays, one fundamental requisite is the abundance of training data, which can be ensured by emerging transportation big data. These models care about the statistical relationship among input features, with little prespecified assumptions. It is the quantity of data that makes machine learning models, which are complex in terms of the number of parameters, effective in predictive tasks. When it comes to future scenario, not in historical data, the trained machine learning model may not be able to present an accurate result, neither do classical parametric models. However, as a Bayesian model, GP can show us the uncertainty of prediction results by providing a posterior distribution instead of merely a point estimate.

B. Incorporating Noises in Observation

To estimate the dynamic road capacity, it is necessary to address noise in data. Despite the increased dimensionality of the features input into the model, we cannot account for all uncertainties in real-world traffic flow, e.g., changes in driving behaviours among drivers and in different time instants. This intrinsic noise in the observed volumes in the dataset may lead to overfitting of the traffic flow model and an inaccurate estimate of capacity. Therefore, in this section, we show how to incorporate noise into the observations in the training set.

A standard GP regression model assumes that the observations are noise-free, i.e., $\mathbf{y} = f(\mathbf{x})$. For aggregated traffic variables, it is more reasonable to incorporate the stochasticity of traffic flow by including a noise term $\boldsymbol{\varepsilon}$, that is, $\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\varepsilon}$. We allow the noise to be included in the GP regression model by adding a constant to the diagonal of the original kernel matrix $\mathbf{K}$. Then, the kernel matrix will be

$$\mathbf{K}_{noisy} = \mathbf{K} + \sigma_K^2 \mathbf{I} \tag{20}$$

where $\mathbf{I}$ denotes the identity matrix, and σ_K^2 denotes the noise

constant. This formulation allows the kernel to incorporate the noise inherent in each sample of the dataset; otherwise, the posterior variance at observed points will be zero. In this case, we can write the joint prior distribution of $\mathbf{Y}$ and Y_* of a zero-mean GP as follows:

$$\begin{pmatrix} \mathbf{Y} \\ Y_* \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{K} + \sigma_K^2 \mathbf{I} & \boldsymbol{\kappa}_* \\ \boldsymbol{\kappa}_*^{\mathrm{T}} & \kappa_{**} \end{pmatrix} \right). \tag{21}$$

The posterior of this GP regression can be written as

$$\mathbf{Y}_* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}_* \sim \mathsf{N}\left(\boldsymbol{\kappa}_*^{\mathrm{T}} \left(\mathbf{K} + \sigma_K^2 \mathbf{I} \right)^{-1} \mathbf{y}, \kappa_{**} - \boldsymbol{\kappa}_*^{\mathrm{T}} \left(\mathbf{K} + \sigma_K^2 \mathbf{I} \right)^{-1} \boldsymbol{\kappa}_* \right) \tag{22}$$

The mean and variance of the posterior are identical to those without noise, except for the kernel matrix inversion term. Accordingly, the solution to a GP without noise is still applicable when observation noise is incorporated. The optimal value of σ_K can be determined by minimising cross-validation errors.

VI. Simulation-Based Experiments

Typical demand-side and supply-side data in reality can usually be obtained in simulation and their influences on traffic flow can be reproduced. Collecting data in simulation can avoid errors in data acquisition; hence there is no need for data preprocessing, which ensures the reliability of the data for model analysis. Although simulation cannot perfectly reflect the actual traffic dynamics, the performance of the proposed model can still be examined. Also note that the proposed model is designed to be versatile to various data sources rather than being restricted to a specific set of input features, and its effectiveness can be validated even when the data used in the simulation are not completely identical to those in reality. Therefore, in this paper, simulations are performed on Vissim to generate a set of data to evaluate the proposed data-driven traffic flow model. In the experiment, traffic volume is used as the label when computing the prediction error of the model, and the dataset is randomly partitioned into two subsets to evaluate the generalisation performance of the model. Two common metrics for the goodness-of-fit of regression, namely mean absolute error (MAE) and mean absolute percentage error (MAPE), are used to evaluate the performance. The GP model used in the experiment adopts the anisotropic Gaussian kernel. The kernel hyperparameters are optimised using the modified Newton method. As one application of the data-driven traffic flow model, we maximise the traffic volume under the given supply-side features to obtain the road capacity. The estimated result is then compared with the maximum observed volume in the simulation.

A. Data Description

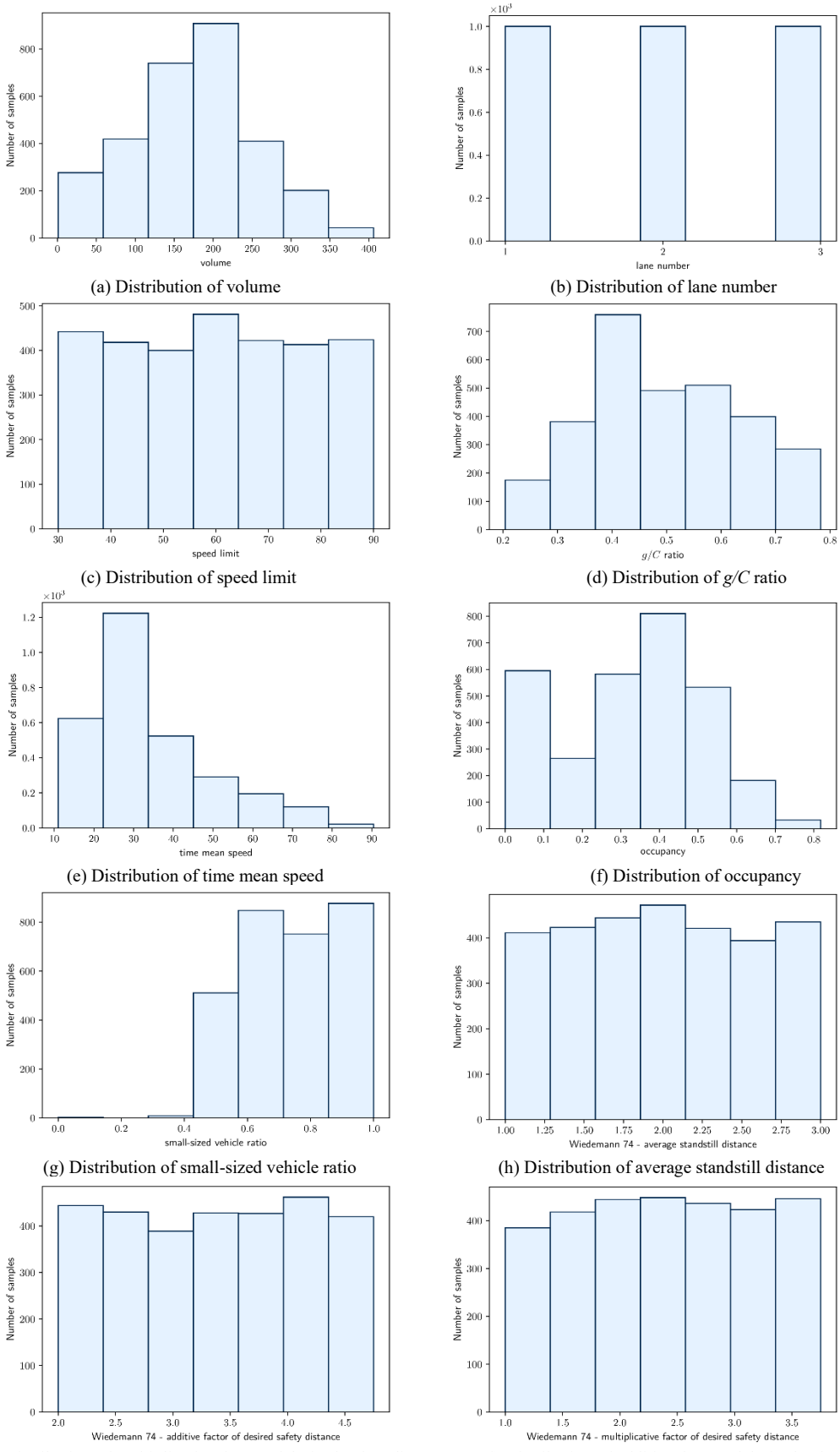
A minimal urban road scenario is built in the microscopic traffic simulator Vissim, containing a basic road section and traffic signals upstream. Three scenarios from one to three lanes are constructed, and 1,000 simulation runs are performed in each scenario. One simulation run lasts one hour, with randomly generated parameters. To show the capability of the proposed model in modelling macroscopic relationships among various traffic-related factors, the simulation setting in this experiment is simplified; for the demonstration, only the signal timing, lane number and speed limit are altered in terms of supply-side variables, and vehicle turning at the intersection is not considered. In this simulation, the signal cycle consists of two phases, one allowing all vehicles on the road to pass, and the other stopping them. The durations of the two phases are independently randomised between 10 and 40 seconds; accordingly, the cycle time ranges from 20 seconds to 80 seconds.

Traffic data are collected by virtual sensors located 200 meters downstream of the traffic signal. The sensors in our simulation mimic both the loop detectors and the LPR system in the real world, which record the instantaneous vehicle speed, vehicle type, and corresponding timestamp when the vehicle enters and leaves the detection area. These data are aggregated over a time window of 15 minutes every time a vehicle is detected, and a small dataset is then constructed by randomly selecting 3,000 samples from the full dataset to reduce the computational burden in the following experiments. In addition, four parameters of Vissim relating to driving behaviour are included in the feature set, namely the three parameters of the Wiedemann 74 car following model and the safety reduction factor of lane changing [65]. Although these data cannot be directly observed in reality, they are added as surrogates for other behavioural data collected through field surveys and video analysis. The distribution of volume and all candidate features are presented in Fig. 9.

Before discussing the performance of the proposed model, it is worth showing the scatter between q and the product of v_{fms} and o (see Fig. 10). The theoretical linear relationship cannot be observed from the Figure, indicating that the traffic deviates from the ideal equilibrium state. This finding consolidates the need to use a data-driven model in real-time traffic management, which is potentially more suitable for modelling complex environments.

B. Model Implementation and Experiment Results

To evaluate the performance of the proposed model, we randomly partition the dataset into two subsets, 80% of which are used as the training set and the other 20% as the test set. MAE



(a) Distribution of volume

(b) Distribution of lane number

(c) Distribution of speed limit

(d) Distribution of g/C ratio

(e) Distribution of time mean speed

(f) Distribution of occupancy

(g) Distribution of small-sized vehicle ratio

(h) Distribution of average standstill distance

(i) Distribution of multiplicative factor of desired safety distance

(j) Distribution of additive factor of desired safety distance

Fig. 9. Distribution of volume and candidate features extracted from the simulation

TABLE I
PERFORMANCE OF VARIOUS DEMAND-SIDE FEATURE COMBINATIONS

Number of Features	occ.	mean speed	small vehicle	W74 stand.	W74 add.	W74 multi.	safety reduc.	**Train MAE**	**Train MAPE**	**Test MAE**	**Test MAPE**
	●							**65.73**	**9.33%**	**71.43**	**10.51%**
		●						71.76	10.87%	77.36	12.39%
			●					101.95	17.66%	110.24	19.49%
1					●			103.01	17.77%	111.12	19.74%
				●				105.14	18.09%	115.57	20.27%
						●		106.08	18.16%	115.53	20.29%
							●	106.07	18.15%	115.65	20.29%
	●		●					**51.61**	**7.37%**	**53.93**	**7.83%**
	●	●						61.35	8.50%	64.69	9.08%
	●			●				58.25	8.42%	61.05	9.20%
	●				●			55.73	8.14%	60.31	9.20%
		●	●					59.32	9.04%	64.80	10.07%
2					●	●		102.53	17.70%	110.73	19.71%
					●		●	102.93	17.76%	111.13	19.74%
				●		●		104.80	18.04%	115.23	20.22%
				●			●	104.52	18.00%	115.45	20.24%
						●	●	105.60	18.09%	115.31	20.24%
	●		●		●			**41.05**	**6.07%**	**42.59**	**6.46%**
	●	●	●					44.68	6.35%	46.31	6.61%
	●		●	●				45.85	6.58%	47.16	6.96%
	●		●			●		50.40	7.25%	52.76	7.70%
	●		●				●	51.29	7.34%	53.74	7.81%
3			●			●	●	101.07	17.57%	109.24	19.40%
				●	●	●		100.98	17.57%	110.07	19.61%
				●	●		●	101.34	17.62%	110.47	19.64%
					●	●	●	102.49	17.70%	110.89	19.75%
				●		●	●	104.15	17.95%	115.06	20.18%
	●	●	●		●			34.44	**5.10%**	**35.00**	**5.25%**
	●		●	●	●			**34.07**	5.21%	35.15	5.56%
	●		●		●	●		37.41	5.67%	39.04	6.08%
	●	●	●	●				41.46	5.86%	43.11	6.17%
	●	●	●			●		43.37	6.20%	44.93	6.47%
4			●	●	●	●		95.36	16.83%	101.65	18.32%
			●	●	●		●	95.92	16.91%	102.23	18.36%
			●		●	●	●	96.64	16.99%	102.03	18.40%
			●	●		●	●	99.10	17.30%	107.56	19.13%
				●	●	●	●	100.93	17.57%	110.04	19.61%
	●	●	●	●	●			**29.69**	**4.48%**	30.80	**4.71%**
	●	●	●		●	●		31.04	4.70%	**30.75**	4.77%
	●		●	●	●	●		29.00	4.65%	29.65	4.96%
	●	●	●		●		●	34.50	5.12%	35.06	5.27%
	●		●	●	●		●	34.11	5.23%	35.29	5.59%
5	●	●		●		●	●	54.57	7.67%	56.51	8.13%
		●	●		●	●	●	44.95	7.30%	48.73	8.17%
		●	●	●		●	●	52.53	8.17%	56.17	8.98%
		●		●	●	●	●	59.73	9.47%	64.26	10.95%
			●	●	●	●	●	95.39	16.84%	101.63	18.31%
	●	●	●	●	●	●		**25.31**	**3.96%**	**25.66**	**4.11%**
	●	●	●	●	●		●	29.79	4.50%	30.94	4.74%
	●	●	●		●	●	●	31.04	4.71%	30.74	4.77%
6	●		●	●	●	●	●	29.03	4.65%	29.72	4.97%
	●	●	●	●		●	●	39.91	5.68%	41.46	5.98%
	●	●		●	●	●	●	42.55	6.19%	45.00	6.85%
		●	●	●	●	●	●	37.63	6.43%	40.64	7.19%
7	●	●	●	●	●	●	●	**25.39**	**3.98%**	**25.73**	**4.12%**

Note: occ. represents occupancy; *small vehicle* represents small vehicle ratio; *W74 stand., W74 add.* and *W74mult.* represent the three parameters of the Wiedemann 74 car following model; *safety reduc*. represents the safety reduction factor of lane changing.

and MAPE are used as the performance metrics, which are formulated as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \tag{23}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% . \tag{24}$$

The anisotropic Gaussian kernel, which can give a smooth posterior estimate and allow for the contributions of various in experiments. The optimal length scales are selected by maximising the LML of the GP regression, where the modified Newton method is used to reduce the probability of the optimisation algorithm getting stuck in the plateau of the hyperparameter surface.

1) Feature Combination Selection

In feature combination selection, we mainly focus on demand-side features. Hence, the three road attribute features, i.e., lane number, *g/C* ratio and speed limit, are always included in the feature set. All combinations of candidate demand-side features are added to the feature set and evaluated. The noise constant σ_K^2 is set to 0.01, the tuning process of which is shown in Appendix C. The performances of these feature combinations are listed in Table I. Due to limited space for demonstration, only the combinations ranking in the top five and bottom five are shown for each group with the same number of features.

A general trend that can be observed from the table is the decreasing estimation error as the number of input features increases. The test MAPE drops by approximately 2.7% for the best combination when the number of features increases from 1 to 2. As the number of features increases, the performance of the model shows only marginal improvement, and the test MAPE reaches 4.12% if all demand-side features are added to the model, which is almost the same as the performance without the last feature, i.e., safety reduction factor of lane changing.

Among all demand-side features, occupancy is the one appearing most frequently in the top five feature sets. Regarding the performance when only one demand-side feature is used, both occupancy and time mean speed, two important features in traffic flow analysis, perform well. As for the other five features, their estimation errors on the test set are more than twice those of the previous two when no other demand-side features are included.

However, combining occupancy and time mean speed does not produce the best result compared with other combinations. The MAPE of the combination of occupancy and small-sized vehicle ratio reaches 7.83%, which is slightly better than the combination of occupancy and time mean speed. As to the combinations of car following parameters, which cannot be directly observed in reality, they have much higher errors than the first three features.

By fixing the lane number and speed limit, we can plot the relationship between volume and occupancy against different *g/C* ratios. As plotted in Fig. 11 using the model with one demand-side feature, i.e., occupancy, various patterns of volume–occupancy curves can be observed. The curve gets flatter as the *g/C* ratio decreases, reflecting its effect on road capacity. Using

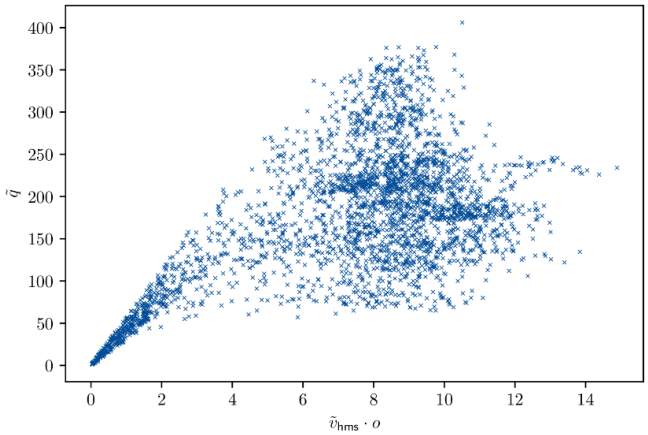


Fig. 10. Scatter plot of $\tilde{v}_{\text{hms}} \cdot o$ and $\tilde{q}$

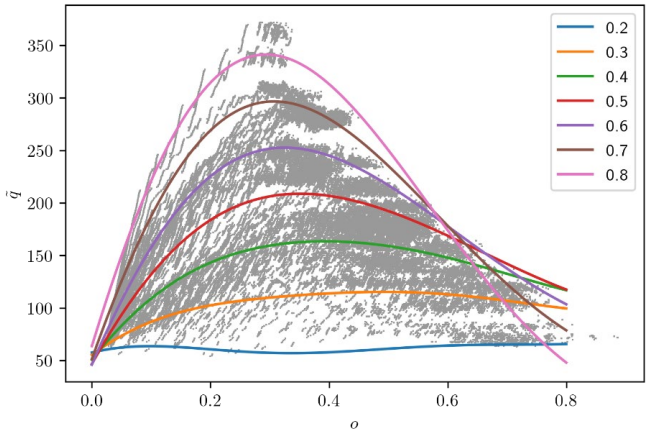


Fig. 11. Volume-occupancy relationship given different g/C ratios

classical FD models, only one curve can be estimated, which will suffer from the dispersion caused by multiple or adaptive signal controlling schemes. In contrast, different volume–occupancy curves can be simultaneously modelled in one model because our model accounts for the *g/C* ratio.

2) Comparison with FD Models

The performances of three classical FD models – the Greenshields model, triangle model and trapezoidal model – are also evaluated. The flow–density relationship described by these models can be formulated as follows.

a) Greenshields FD

$$q = \rho v_f \left(1 - \rho / \rho_j\right) \tag{25}$$

where v_f and ρ_j are the two parameters of this functional form.

b) Triangle FD

$$q = \begin{cases} v_f \rho, & \rho \le \rho_c \\ v_f \rho_c (\rho - \rho_j)/(\rho_c - \rho_j), & \rho > \rho_c \end{cases} \tag{26}$$

where v_f, ρ_j, and ρ_c are the three parameters of this functional form.

TABLE II
PERFORMANCE OF FD MODELS

Supply-side features		Greenshields		Triangle		Trapezoidal		Our model	
Lane number	Speed limit	Test MAE	Test MAPE	Test MAE	Test MAPE	Test MAE	Test MAPE	Test MAE	Test MAPE
1	40	169.86	28.55%	155.59	26.37%	153.72	26.06%	50.70	7.50%
1	50	121.83	18.05%	100.84	15.37%	97.72	14.49%	49.53	7.36%
1	60	178.77	19.74%	160.40	17.96%	149.68	17.00%	71.14	8.01%
2	40	92.85	13.35%	93.99	13.78%	97.17	14.17%	55.39	7.58%
2	50	127.39	19.09%	118.89	18.77%	115.87	18.33%	69.89	11.28%
2	60	205.57	36.57%	113.71	17.32%	110.80	16.62%	55.17	8.58%
3	40	118.27	19.80%	120.69	19.73%	123.02	20.18%	56.13	8.97%
3	50	204.03	31.91%	173.56	26.76%	183.77	27.70%	52.79	7.54%
3	60	149.70	28.06%	124.46	21.31%	121.68	18.85%	85.72	18.05%

TABLE III
COMPARISON BETWEEN DATA-DRIVEN MODELS

	Test MAE	Test MAPE
SVM	159.03	24.15%
DT	39.58	6.29%
ANN	222.88	38.11%
Our model	**31.72**	**4.77%**

TABLE IV
LENGTH SCALES OF DIFFERENT FEATURES

Features	*g/C* ratio	lane number	speed limit	occ.	mean speed	small vehicle ratio	W74 stand.	W74 add.	W74 multi.	safety reduc.
Length scale	6.55	100	100	2.06	14.5	24.9	94	46	100	100

c) Trapezoidal FD

$$q=\begin{cases} v_f\rho, & \rho \le \rho_{c1} \\ v_f\rho_{c1}, & \rho_{c1} \le \rho < \rho_{c2} \\ v_f\rho_{c2}(\rho-\rho_j)/(\rho_{c2}-\rho_j), & \rho > \rho_{c2} \end{cases} \tag{27}$$

where v_f, ρ_j, ρ_{c1}, and ρ_{c2} are the four parameters of this functional form.

Because classical FD models cannot model traffic on distinct roads, we split the data by lane number and speed limit. The results of FD models on data with speed limits of 40 km/h, 50 km/h and 60 km/h are shown in Table II. It is clear that the FD models are outperformed by the proposed data-driven model. In comparison, the differences between the three FD models are minor. As discussed, traffic flow is inherently complex, especially on urban roads. Although physical models provide powerful insights into various traffic phenomena, they may be unsuitable for real-time urban traffic management.

3) Comparison with data-driven models

We now benchmark the performance of the proposed model to that of three widely used data-driven methods: the support vector machine (SVM), the decision tree (DT), and the artificial neural network (ANN). In training these models, all the 10 features are considered simultaneously.

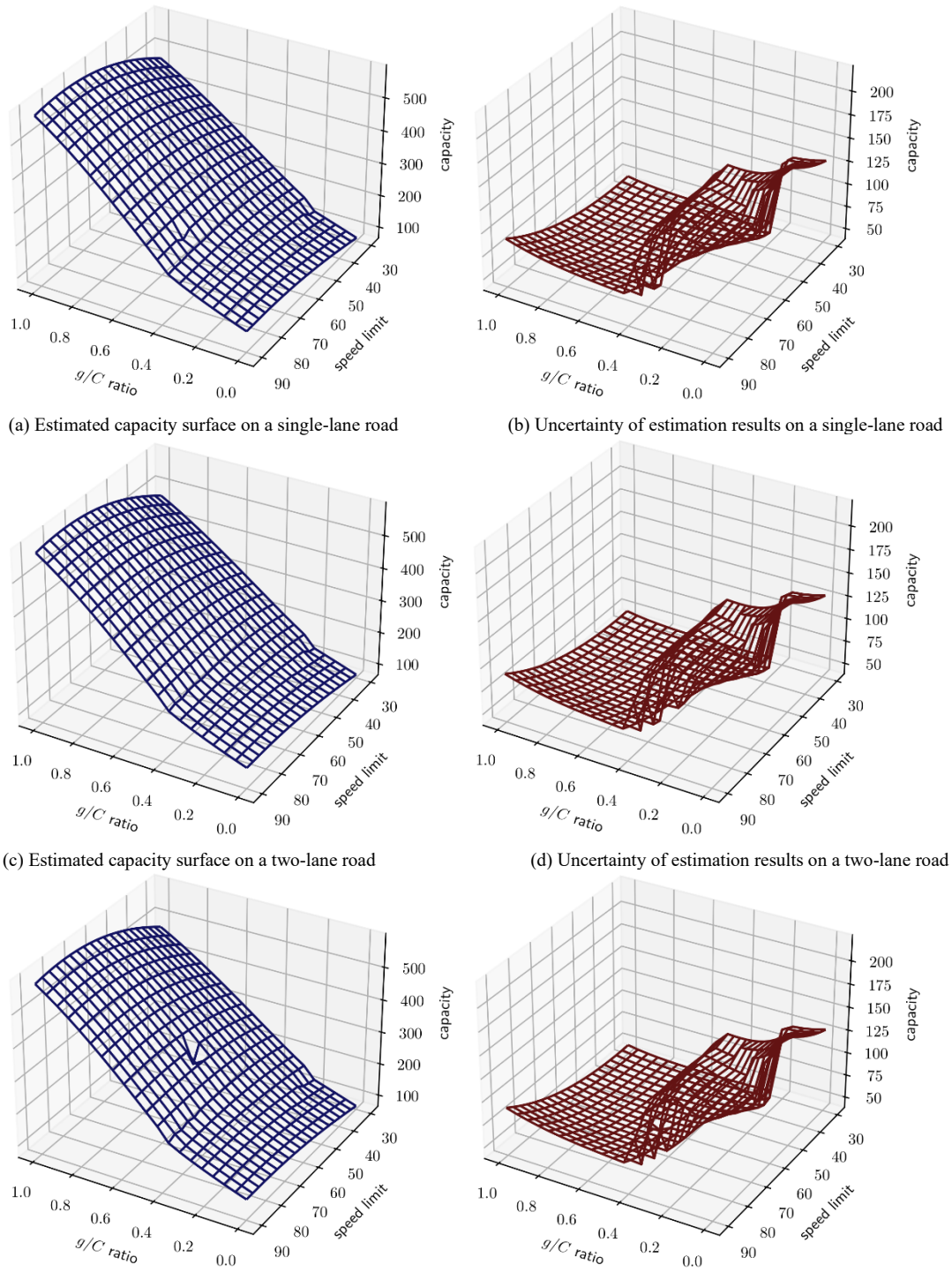
Experimental results are shown in Table III. Results indicate that the proposed model outperforms the benchmark methods. The decision tree model has a similar performance to the proposed model. Whereas, the reason why GP is chosen is mainly influenced by the requirement of capacity estimation. Recall that the capacity is calculated by:

$$c(\boldsymbol{x}_s)=\arg\max_{\boldsymbol{x}_d} f_{ds}(\boldsymbol{x}_d,\boldsymbol{x}_s) \tag{28}$$

For a given supply condition, the optimisation problem above is implicitly constrained for $\boldsymbol{x}_d \in \boldsymbol{X}_d$. For example, it is impossible that the traffic on a narrow single-lane road is crowded in terms of density and fast-moving in terms of speed at the same time. GP is a Bayesian model, where a prior mean function is used to encode our prior knowledge on the values of prediction labels. The prior mean ensures that the posterior mean of infeasible $\boldsymbol{x}_d \in \boldsymbol{X}_d$, which is not observed in the training data, stays reasonable. In this way, they will not influence the capacity estimation. Nevertheless, other models such as SVM, DT and ANN fail to provide reasonable estimations under all the cases, as they cannot infer the feasible region $\boldsymbol{X}_d$ from the data; the estimated capacity is usually unreasonably high.

We adopt the length scales of the estimated GP kernel as an indicator of feature importance. An important feature often results in frequent and dramatic gradient change in the posterior mean on that dimension. The rationale behind the relationship between length scale and feature importance is explained in Appendix D. The length scales corresponding to different features are listed in Table IV. Clearly, occupancy, *g/C* ratio, time mean speed and small-sized vehicle ratio are the four features that contribute most to the estimation.

By optimising the demand-side features, we can obtain the capacity given specific supply-side features. According to Table IV, we can infer that the capacity (vehicles per hour per lane) does not differ much with respect to lane number and speed limit. We validate this conjecture by plotting the capacity estimation surface against these features in Fig. 12 (a), (c) and (e). Note that the safety reduction factor of lane changing is excluded from the feature set when plotting due to the minor improvement in estimation accuracy, as shown in Table I. No significant differences can be observed among the three plots, and



(a) Estimated capacity surface on a single-lane road

(b) Uncertainty of estimation results on a single-lane road

(c) Estimated capacity surface on a two-lane road

(d) Uncertainty of estimation results on a two-lane road

(e) Estimated capacity surface on a three-lane road

(f) Uncertainty of estimation results on a three-lane road

Fig. 12. *Estimated capacity surface*

the derivatives along the axis of the speed limit are almost flat. The standard deviations of the estimation results are given in Fig. 12 (b), (d) and (f). It can be observed that a larger variance appears for a small or large *g/C* ratio, when few observations are included in the dataset (see Fig. 12 (d)).

VII. Conclusion

FD-based traffic flow models have long been the primary tool for modelling traffic states and have provided the foundations for subsequent policymaking and management practices. Numerous studies have focused on the theoretical basis of the FD, in an effort to unveil traffic patterns and explain complex traffic phenomena. Most existing models have been validated and proven effective in modelling highway traffic, but urban traffic modelling is still a challenge due to the volatility of urban driving behaviours. When the FD is used to model urban traffic,

a large dispersion of data can be observed, which cannot be adequately fitted by parametric models. The reasons are two-fold:

1. Due to interference from factors like traffic signals, unmotorised traffic and entrances/exits from road-side constructions, urban traffic inherently involves greater uncertainty.
2. Supply-side factors do not always remain constant. Changes in the weather and in signal plans will have a significant impact on road traffic.

Consequently, classical FD-based traffic models are insufficient to explain variations in traffic variables.

Faced with this issue, we propose a data-driven traffic flow model in our study, which fully utilises traffic big data and is capable of modelling urban traffic states more accurately. The novel formulation of traffic flow incorporates variables relating to both traffic state and road supply. This extends the model so it can adapt to multiple roads in a given region and account for changes in external influencing factors like weather and signal plans. In addition, compared with classical FD-based models, the data-driven model does not require strong assumptions on the functional form of the relationship between traffic variables. Instead, the proposed model is based on the flexible GP regression model, which can capture complicated patterns in traffic dynamics. Moreover, a modified Newton method for GP-based traffic flow model is designed to avoid getting stuck in the plateau of the optimisation objective during the training process of the GP regression model.

Simulation-based experiments are used to evaluate the performance of the proposed model. The Gaussian kernel is adopted to construct the GP regression model used in the experiment, where noise constants are added to the diagonal of the kernel matrix to account for noise in traffic observations. The length scales of the Gaussian kernel are optimised through the modified Newton method. Compared with classical modelling approaches, the proposed model reduces the estimation error of the MAPE from around 20% to less than 10%. Road capacity estimation is one typical application of the data-driven traffic flow model. Based on the specific physical attributes of a road, the traffic volume can be maximised, which gives the capacity. The results indicate that the proposed model is effective and suitable for modelling complex urban traffic flow.

Our study reveals the possibility of including more variables in the traffic flow model and substituting a parametric model for a machine learning model, which yields improved fitting performance. In addition to the performance improvement, concerning the variables in our model, not only (demand-side) traffic state variables, but also supply-side variables are included; it enables us to connect the physical attributes of roads to the traffic states on them. This is valuable in allowing the estimation of dynamic road capacity. In regard to the model specification testing, the feature combination selection and the estimated length scales w.r.t features together show that among all can candidate features, occupancy, time mean speed, small-sized vehicle ratio, and *g/C* ratio contribute the most to the estimation of traffic volume, where the latter two are seldom incorporated by previous research.

For future research, improvements can be made in two areas: the model for volume estimation and the use of multi-source data. Regarding the model, GP regression can be improved to support efficient computation with large-scale data. The training of an ordinary GP regression model is constrained by the computation of the inversion of the kernel matrix, whose time complexity is $\mathrm{O}(N^3)$, where N is the number of samples [66]. In addition, there is plenty of room to improve the optimisation algorithm for GP regression. Finding an efficient way to jump over the plateau of the optimisation objective surface is a challenging but critical issue. Another meaningful extension is to develop data-driven traffic flow models considering the case of actuated and adaptive control, under which the green/cycle (g/C) ratio varies with traffic flows. Aside from the model, the input features can be refined by searching for variables that can better represent the traffic state. Multi-source data, including floating car data and cellphone data, can be introduced to extend the information fed into the model.

REFERENCES

[1] B. D. Greenshields, "A study of capacity," in *Proceedings of the 14th Annual Meeting of the Highway Research Board*, Washington, D.C., 1934, vol. 14, pp. 448–477.

[2] M. J. Lighthill and G. B. Whitham, "On kinematic waves II. A theory of traffic flow on long crowded roads," *Proc. R. Soc. Lond. Ser. Math. Phys. Sci.*, vol. 229, no. 1178, pp. 317–345, 1955, doi: 10.1098/rspa.1955.0089.

[3] P. I. Richards, "Shock waves on the highway," *Oper. Res.*, vol. 4, no. 1, pp. 42–51, 1956, doi: 10.1287/opre.4.1.42.

[4] C. F. Daganzo, *Fundamentals of Transportation and Traffic Operations*. Emerald Group Publishing Limited, 1997. doi: 10.1108/9780585475301.

[5] M. Treiber and A. Kesting, *Traffic Flow Dynamics*. Berlin, Heidelberg: Springer, 2013. doi: 10.1007/978-3-642-32460-4.

[6] M. J. Cassidy, "Bivariate relations in nearly stationary highway traffic," *Transp. Res. Part B Methodol.*, vol. 32, no. 1, pp. 49–59, 1998, doi: 10.1016/S0191-2615(97)00012-X.

[7] B. Coifman, "Empirical flow-density and speed-spacing relationships: Evidence of vehicle length dependency," *Transp. Res. Part B Methodol.*, vol. 78, pp. 54–65, 2015, doi: 10.1016/j.trb.2015.04.006.

[8] X. Wu, H. X. Liu, and N. Geroliminis, "An empirical analysis on the arterial fundamental diagram," *Transp. Res. Part B Methodol.*, vol. 45, no. 1, pp. 255–266, 2011, doi: 10.1016/j.trb.2010.06.003.

[9] N. Geroliminis and C. F. Daganzo, "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings," *Transp. Res. Part B Methodol.*, vol. 42, no. 9, pp. 759–770, 2008, doi: 10.1016/j.trb.2008.02.002.

[10] H. M. Zhang, "A non-equilibrium traffic model devoid of gas-like behavior," *Transp. Res. Part B Methodol.*, vol. 36, no. 3, pp. 275–290, 2002, doi: 10.1016/S0191-2615(00)00050-3.

[11] H. M. Zhang, "A theory of nonequilibrium traffic flow," *Transp. Res. Part B Methodol.*, vol. 32, no. 7, pp. 485–498, 1998, doi: 10.1016/S0191-2615(98)00014-9.

[12] L. C. Edie, "Discussion of traffic stream measurements and definitions," in *Proceedings of International Symposium on the Theory of Traffic Flow*, Paris, France, 1965, pp. 139–154.

[13] V. Knoop, S. P. Hoogendoorn, and H. van Zuylen, "Empirical differences between time mean speed and space mean speed," in *Traffic and Granular Flow '07*, Berlin, Heidelberg, 2009, pp. 351–356. doi: 10.1007/978-3-540-77074-9_36.

[14] W.-L. Jin and H. M. Zhang, "The inhomogeneous kinematic wave traffic flow model as a resonant nonlinear system,"

Transp. Sci., vol. 37, no. 3, pp. 294–311, 2003, doi: 10.1287/trsc.37.3.294.16046.

[15] M. van Aerde and H. Rakha, "Multivariate calibration of single regime speed-flow-density relationships," in *Vehicle Navigation and Information Systems Conference Proceedings*, Seattle, WA, USA, 1995, pp. 334–341. doi: 10.1109/VNIS.1995.518858.

[16] M. Papageorgiou, H. Hadj-Salem, and J.-M. Blosseville, "ALINEA: A Local Feedback Control Law for On-Ramp Metering," *Transp. Res. Rec.*, vol. 1320, pp. 58–64, 1991.

[17] M. Papageorgiou, E. Kosmatopoulos, and I. Papamichail, "Effects of Variable Speed Limits on Motorway Traffic Flow," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2047, no. 1, pp. 37–48, 2008, doi: 10.3141/2047-05.

[18] H. J. Payne, "Models of freeway traffic and control," *Math. Models Public Syst.*, vol. 1, no. 1, pp. 51–61, 1971.

[19] A. Aw and M. Rascle, "Resurrection of 'second order' models of traffic flow," *SIAM J. Appl. Math.*, vol. 60, no. 3, pp. 916–938, 2000, doi: 10.1137/S0036139997332099.

[20] G. C. K. Wong and S. C. Wong, "A multi-class traffic flow model – an extension of LWR model with heterogeneous drivers," *Transp. Res. Part Policy Pract.*, vol. 36, no. 9, pp. 827–841, 2002, doi: 10.1016/S0965-8564(01)00042-8.

[21] G. F. Newell, "Nonlinear effects in the dynamics of car following," *Oper. Res.*, vol. 9, no. 2, pp. 209–229, 1961, doi: 10.1287/opre.9.2.209.

[22] V. L. Knoop and W. Daamen, "Automatic fitting procedure for the fundamental diagram," *Transp. B Transp. Dyn.*, vol. 5, no. 2, pp. 129–144, 2017, doi: 10.1080/21680566.2016.1256239.

[23] C. F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transp. Res. Part B Methodol.*, vol. 28, no. 4, pp. 269–287, 1994, doi: 10.1016/0191-2615(94)90002-7.

[24] N. Wu, "A new approach for modeling of fundamental diagrams," *Transp. Res. Part Policy Pract.*, vol. 36, no. 10, pp. 867–883, 2002.

[25] J. Li and H. M. Zhang, "Fundamental diagram of traffic flow: new identification scheme and further evidence from empirical data," *Transp. Res. Rec.*, vol. 2260, no. 1, pp. 50–59, 2011, doi: 10.3141/2260-06.

[26] X. Qu, S. Wang, and J. Zhang, "On the fundamental diagram for freeway traffic: A novel calibration approach for single-regime models," *Transp. Res. Part B Methodol.*, vol. 73, pp. 91–102, 2015, doi: 10.1016/j.trb.2015.01.001.

[27] G. Dervisoglu, G. Gomes, J. Kwon, R. Horowitz, and P. Varaiya, "Automatic calibration of the fundamental diagram and empirical observations on capacity," in *Transportation Research Board 88th Annual Meeting*, Washington, D.C., 2009.

[28] T. Seo, Y. Kawasaki, T. Kusakabe, and Y. Asakura, "Fundamental diagram estimation by using trajectories of probe vehicles," *Transp. Res. Part B Methodol.*, vol. 122, pp. 40–56, 2019, doi: 10.1016/j.trb.2019.02.005.

[29] M. S. Ahmed and A. R. Cook, "Analysis of freeway traffic time-series data by using Box–Jenkins techniques," *Transp. Res. Rec.*, no. 722, pp. 1–9, 1979.

[30] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, 2014, doi: 10.1109/TITS.2014.2345663.

[31] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. Part C Emerg. Technol.*, vol. 79, pp. 1–17, 2017, doi: 10.1016/j.trc.2017.02.024.

[32] C. Meng, X. Yi, L. Su, J. Gao, and Y. Zheng, "City-wide traffic volume inference with loop detector data and taxi trajectories," in *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Redondo Beach, CA, USA, 2017, pp. 1–10. doi: 10.1145/3139958.3139984.

[33] X. Wang, R. Jiang, L. Li, Y. Lin, X. Zheng, and F.-Y. Wang, "Capturing car-following behaviors by deep learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 910–920, 2018, doi: 10.1109/TITS.2017.2706963.

[34] S. Sun, R. Huang, and Y. Gao, "Network-scale traffic modeling and forecasting with graphical lasso and neural networks," *J. Transp. Eng.*, vol. 138, no. 11, pp. 1358–1367, 2012, doi: 10.1061/(ASCE)TE.1943-5436.0000435.

[35] L. Lin, J. Li, F. Chen, J. Ye, and J. Huai, "Road traffic speed prediction: A probabilistic model fusing multi-source data," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 7, pp. 1310–1323, 2018, doi: 10.1109/TKDE.2017.2718525.

[36] Y. Yuan, Z. Zhang, X. T. Yang, and S. Zhe, "Macroscopic traffic flow modeling with physics regularized Gaussian process: A new insight into machine learning applications in transportation," *Transp. Res. Part B Methodol.*, vol. 146, pp. 88–110, 2021, doi: 10.1016/j.trb.2021.02.007.

[37] J. Li, J. Boonaert, A. Doniec, and G. Lozenguez, "Multi-models machine learning methods for traffic flow estimation from Floating Car Data," *Transp. Res. Part C Emerg. Technol.*, vol. 132, p. 103389, 2021, doi: 10.1016/j.trc.2021.103389.

[38] M. Sederlin, X. Ma, and J. Jin, "A Hybrid Modelling Approach for Traffic State Estimation at Signalized Intersections," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, Indianapolis, IN, USA, 2021, pp. 3604–3609. doi: 10.1109/ITSC48978.2021.9564540.

[39] F. L. Hall, "Traffic Stream Characteristics," in *Traffic Flow Theory: A State-of-the-Art Report (Revised)*, US Federal Highway Administration, 2001.

[40] J. G. Wardrop, "Some theoretical aspects of road traffic research," *Proc. Inst. Civ. Eng.*, vol. 1, no. 3, pp. 325–362, 1952, doi: 10.1680/ipeds.1952.11259.

[41] X. Qu, J. Zhang, and S. Wang, "On the stochastic fundamental diagram for freeway traffic: Model development, analytical properties, validation, and extensive applications," *Transp. Res. Part B Methodol.*, vol. 104, pp. 256–271, 2017, doi: 10.1016/j.trb.2017.07.003.

[42] H. M. Zhang, "A mathematical theory of traffic hysteresis," *Transp. Res. Part B Methodol.*, vol. 33, no. 1, pp. 1–23, 1999, doi: 10.1016/S0191-2615(98)00022-8.

[43] H. Wang, D. Ni, Q.-Y. Chen, and J. Li, "Stochastic modeling of the equilibrium speed-density relationship," *J. Adv. Transp.*, vol. 47, no. 1, pp. 126–150, 2013, doi: 10.1002/atr.172.

[44] B. S. Kerner, "Three-phase traffic theory and highway capacity," *Phys. Stat. Mech. Its Appl.*, vol. 333, pp. 379–440, 2004, doi: 10.1016/j.physa.2003.10.017.

[45] T. Q. Tang, H. J. Huang, S. G. Zhao, and H. Y. Shang, "A new dynamic model for heterogeneous traffic flow," *Phys. Lett. A*, vol. 373, no. 29, pp. 2461–2466, 2009, doi: 10.1016/j.physleta.2009.05.006.

[46] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[47] M. G. Genton, "Classes of kernels for machine learning: a statistics perspective," *J. Mach. Learn. Res.*, vol. 2, pp. 299–312, 2001.

[48] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. New York, NY: Springer New York, 1996. doi: 10.1007/978-1-4612-0745-0.

[49] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[50] M. M. Minderhoud, H. Botma, and P. H. Bovy, "Assessment of roadway capacity estimation methods," *Transportation research record,* vol. 1572, no. 1, pp. 59-67, 1997.

[51] H. C. Manual, "HCM2010," Transportation Research Board, National Research Council, Washington, DC, vol. 1207, 2010.
[52] G. Dervisoglu, G. Gomes, J. Kwon, R. Horowitz, and P. Varaiya, "Automatic calibration of the fundamental diagram and empirical observations on capacity," in *Transportation Research Board 88th Annual Meeting*, 2009, vol. 15: Citeseer, pp. 31-59.
[53] R. Chauhan, P. Kumar, S. Arkatkar, A. Dhamaniya, and P. K. Sahu, "Examining deterministic and probabilistic capacity estimation methods under mixed traffic using empirical data," *Case Studies on Transport Policy,* vol. 9, no. 4, pp. 1888-1899, 2021.
[54] X. Shi and X. Li, "Constructing a fundamental diagram for traffic flow with automated vehicles: Methodology and demonstration," *Transportation Research Part B: Methodological,* vol. 150, pp. 279-292, 2021.
[55] X. Z. Simon et al., "A meso-to-macro cross-resolution performance approach for connecting polynomial arrival queue model to volume-delay function with inflow demand-to-capacity ratio," *Multimodal Transportation*, vol. 1, no. 2, p. 100017, 2022.
[56] M. Chang and Y. Kim, "Development of capacity estimation method from statistical distribution of observed traffic flow," in *Proceedings: Fourth International Symposium on Highway Capacity*, 2000, pp. 299-309.
[57] M. J. Cassidy and R. L. Bertini, "Some traffic features at freeway bottlenecks," *Transportation Research Part B: Methodological,* vol. 33, no. 1, pp. 25-42, 1999.
[58] S. Heshami, L. Kattan, Z. Gong, and S. Aalami, "Deterministic and stochastic freeway capacity analysis based on weather conditions," *Journal of Transportation Engineering, Part A: Systems,* vol. 145, no. 5, p. 04019016, 2019.
[59] J. Zhu, I. Tasic, and X. Qu, "Flow-level coordination of connected and autonomous vehicles in multilane freeway ramp merging areas," Multimodal Transportation, vol. 1, no. 1, p. 100005, 2022.
[60] R. Bak and M. Kiec, "Influence of midblock pedestrian crossings on urban street capacity," *Transportation research record,* vol. 2316, no. 1, pp. 76-83, 2012.
[61] S. Biswas, S. Chandra, and I. Ghosh, "Side friction parameters and their influences on capacity of Indian undivided urban streets," *International Journal of Transportation Science and Technology,* vol. 10, no. 1, pp. 1-19, 2021.
[62] N. Geroliminis and B. Boyacı, "The effect of variability of urban systems characteristics in the network capacity," *Transportation Research Part B: Methodological,* vol. 46, no. 10, pp. 1607-1623, 2012.
[63] M. Asgharzadeh and A. Kondyli, "Comparison of highway capacity estimation methods," *Transportation Research Record*, vol. 2672, no. 15, pp. 75-84, 2018.
[64] D. Chen and S. Ahn, "Capacity-drop at extended bottlenecks: Merge, diverge, and weave," *Transportation Research Part B*: Methodological, vol. 108, pp. 1-20, 2018.
[65] PTV, *PTV Vissim 2020 User Manual*. Karlsruhe, Germany: PTV Group, 2020.
[66] A. G. Wilson and H. Nickisch, "Kernel interpolation for scalable structured gaussian processes (KISS-GP)," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 2015, vol. 37, pp. 1775–1784. doi: 10.5555/3045118.3045307.
[67] R. Yan and S. Wang, "Integrating prediction with optimization: models and applications in transportation management, *Multimodal Transportation*, 1 (2022), 1–5, 2022.

APPENDIX

A. Notations

See Table AI.

B. Proof of Proposition 1

Proposition 1. For a GP with an anisotropic Gaussian kernel parameterised by a length scale vector $\boldsymbol{\theta} = [l_1, l_2, \ldots, l_m]$, where $l_i > 0$, $i = 1, 2, \ldots, m$, we have

$$\lim_{\boldsymbol{\theta} \to \mathbf{0}^+} \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{29}$$

Proof. Based on (12) and (13), the derivative of a Gaussian kernel with respect to length scale l_k is

$$\begin{aligned}
\frac{\partial \kappa_{ij}}{\partial l_k} &= \frac{\partial}{\partial l_k} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{\Lambda}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\right) \\
&= \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{\Lambda}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\right) \frac{\partial}{\partial l_k}\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{\Lambda}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\right) \\
&= \kappa_{ij} \frac{\partial}{\partial l_k}\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{\Lambda}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\right) \\
&= \kappa_{ij} \frac{\partial}{\partial l_k}\left(-\frac{1}{2}\sum_{s=1}^{m} l_s^{-2}(x_{i,s} - x_{j,s})^2\right) \\
&= \kappa_{ij} \frac{\partial}{\partial l_k}\left(-\frac{1}{2l_k^2}(x_{i,k} - x_{j,k})^2\right) \\
&= \frac{1}{l_k^3} \kappa_{ij} (x_{i,k} - x_{j,k})^2
\end{aligned} \tag{30}$$

where κ_{ij} is used as a shorthand for $\kappa_{\mathrm{G,ani}}(\mathbf{x}_i, \mathbf{x}_j)$, and $x_{i,k}$ denotes the k-th element of $\mathbf{x}_i$. Then, its limit as l_k approaches zero can be derived,

$$\begin{aligned}
\lim_{l_k \to 0^+} \frac{\partial \kappa_{ij}}{\partial l_k} &= \lim_{l_k \to 0^+} \frac{1}{l_k^3} \kappa_{ij} (x_{i,k} - x_{j,k})^2 \\
&= (x_{i,k} - x_{j,k})^2 \lim_{l_k \to 0^+} \frac{1}{l_k^3} \kappa_{ij} \\
&= (x_{i,k} - x_{j,k})^2 \lim_{l_k \to 0^+} \frac{\exp\left(-\frac{1}{2}\sum_{s=1}^{m} l_s^{-2}(x_{i,s} - x_{j,s})^2\right)}{l_k^3}.
\end{aligned} \tag{31}$$

Denote τ by l_k^{-2} and ζ by $-\frac{1}{2}\sum_{s=1, s\neq k}^{m} l_s^{-2}(x_{i,s} - x_{j,s})^2$; it follows that

$$\begin{aligned}
\lim_{l_k \to 0^+} \frac{\partial \kappa_{ij}}{\partial l_k} &= (x_{i,k} - x_{j,k})^2 \lim_{\tau \to +\infty} \frac{\exp\left(\zeta - \frac{1}{2}\tau(x_{i,k} - x_{j,k})^2\right)}{\tau^{-3/2}} \\
&= (x_{i,k} - x_{j,k})^2 \lim_{\tau \to +\infty} \frac{\tau^{3/2}}{\exp\left(\tau(x_{i,k} - x_{j,k})^2/2 - \zeta\right)}.
\end{aligned} \tag{32}$$

As $\lim_{\tau \to +\infty} \tau^{3/2} = +\infty$ and $\lim_{\tau \to +\infty} \exp\left(\tau(x_{i,k} - x_{j,k})^2/2 - \zeta\right) = +\infty$, the following equations can be derived using L'Hôpital's rule.

$$\begin{aligned}
\lim_{l_k \to 0^+} \frac{\partial \kappa_{ij}}{\partial l_k} &= (x_{i,k} - x_{j,k})^2 \lim_{\tau \to +\infty} \frac{\frac{3}{2}\tau^{1/2}}{\exp\left(\frac{1}{2}\tau(x_{i,k} - x_{j,k})^2 - \zeta\right)\frac{1}{2}(x_{i,k} - x_{j,k})^2} \\
&= \lim_{\tau \to +\infty} \frac{3\tau^{1/2}}{\exp\left(\frac{1}{2}\tau(x_{i,k} - x_{j,k})^2 - \zeta\right)}
\end{aligned}. \tag{33}$$

Again, L'Hôpital's rule yields

$$\begin{aligned}
\lim_{l_k \to 0^+} \frac{\partial \kappa_{ij}}{\partial l_k} &= \lim_{\tau \to +\infty} \frac{\frac{3}{2}\tau^{-1/2}}{\exp\left(\frac{1}{2}\tau(x_{i,k} - x_{j,k})^2 - \zeta\right)\frac{1}{2}(x_{i,k} - x_{j,k})^2} \\
&= \lim_{\tau \to +\infty} \frac{3}{\exp\left(\frac{1}{2}\tau(x_{i,k} - x_{j,k})^2 - \zeta\right)\tau^{1/2}(x_{i,k} - x_{j,k})^2} = 0.
\end{aligned} \tag{34}$$

We generalise to all elements of $\mathbf{K}$ and have

$$\lim_{l_k \to 0^+} \frac{\partial \mathbf{K}}{\partial l_k} = \lim_{l_k \to 0^+} \frac{\partial}{\partial l_k} \begin{pmatrix} \kappa_{11} & \kappa_{12} & \cdots & \kappa_{1n} \\ \kappa_{21} & \kappa_{22} & \cdots & \kappa_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_{n1} & \kappa_{n2} & \cdots & \kappa_{nn} \end{pmatrix} = \mathbf{0}. \tag{35}$$

This can be further extended to all length scales, that is, $\lim_{\boldsymbol{\theta} \to \mathbf{0}^+} \frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}} = \mathbf{0}$. According to the definition of an anisotropic Gaussian kernel,

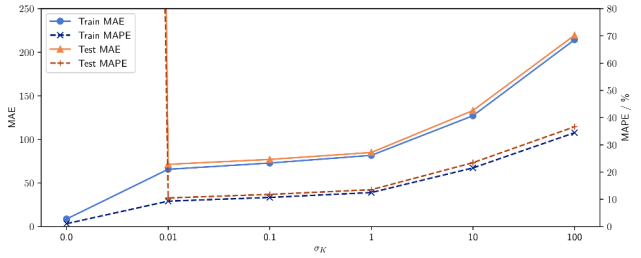
$$\lim_{\boldsymbol{\theta} \to \mathbf{0}^+} \kappa_{ij} = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases} \tag{36}$$

This gives $\lim_{\boldsymbol{\theta} \to \mathbf{0}^+} \mathbf{K} = \mathbf{I}$, and accordingly, $\lim_{\boldsymbol{\theta} \to \mathbf{0}^+} \mathbf{K}^{-1} = \mathbf{I}$.

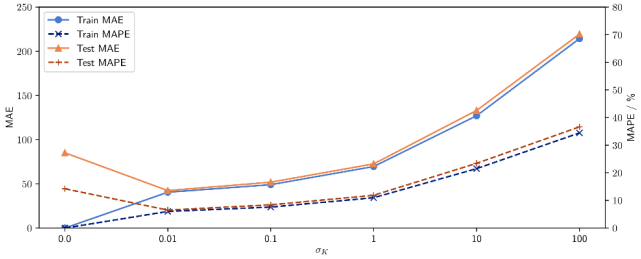
Finally, $\lim_{\boldsymbol{\theta} \to \mathbf{0}^+} \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$ can be proved by means of (17).

TABLE AI
NOTATIONS

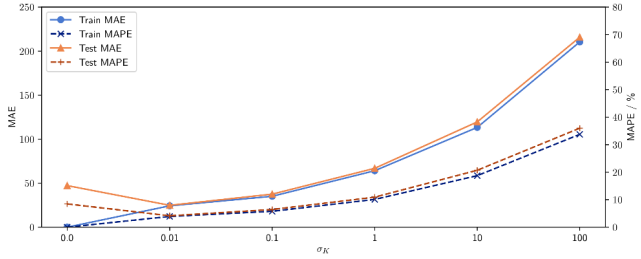
Constants	
T	Duration for aggregating traffic state variables.
L	Length of the road section for observation.
Sets	
V_T	Set of vehicles observed during period T.
V_L	Set of vehicles observed in a road section of length L.
Variables	
q	Flow (defined at a time instant on a long roadway).
ρ	Density (defined at a time instant on a long roadway).
v	Space mean speed (defined at a time instant on a long roadway).
$\tilde{q}$	Volume (defined on a cross section of road for a long time).
$\tilde{\rho}$	Density (defined on a cross section of road for a long time).
$\tilde{v}_{\mathrm{hms}}$	Harmonic mean speed (defined on a cross section of road for a long time).
$\tilde{v}_{\mathrm{tms}}$	Time mean speed (defined on a cross section of road for a long time).
v_i	The instantaneous speed of the i-th vehicle in V_T or V_L.
o_i	The duration that the i-th vehicle in V_T occupying a fixed traffic sensor.



(a) With one demand-side features: occupancy



(b) With four demand-side features: occupancy, small-sized vehicle ratio, Wiedemann 74 average standstill distance, Wiedemann 74 additive factor of desired safety distance



(c) With all candidate demand-side features

Fig. A1. Changes in CV performance against different noise constant

C. *Optimising Noise Constant*

A simple brute search is adopted to determine the optimal noise constant to be added to the Gaussian kernel. The changes in the cross-validation (CV) performance of the test set against different noise constants is plotted in Fig. A1. It can be observed that the error drops steeply and then slowly increases as the noise constant becomes larger. Regardless of the feature set, the optimal value of the noise constant generally falls within the range of $(0, 0.1)$. Therefore, it is reasonable to set $\sigma_K = 0.01$ during feature selection.

D. *Length Scale and Feature Importance*

In this section, we show how the length scale correlates with the feature importance. Consider the derivative of the Gaussian kernel function with respect to an arbitrary input $\mathbf{x}_i$,

$$\begin{aligned}\frac{\partial \kappa_{ij}}{\partial \mathbf{x}_i} &= \frac{\partial}{\partial \mathbf{x}_i}\exp\left(-\frac{1}{2}\left(\mathbf{x}_i-\mathbf{x}_j\right)^{\text{ú}}\mathbf{\Lambda}\left(\mathbf{x}_i-\mathbf{x}_j\right)\right)\\ &= \frac{\partial}{\partial \mathbf{x}_i}\left(-\frac{1}{2}\left(\mathbf{x}_i-\mathbf{x}_j\right)^{\text{ú}}\mathbf{\Lambda}\left(\mathbf{x}_i-\mathbf{x}_j\right)\right)\kappa_{ij}\\ &= -\mathbf{\Lambda}(\mathbf{x}_i-\mathbf{x}_j)\kappa_{ij}.\end{aligned} \tag{37}$$

When dealing with a specific feature dimension, the derivative will be

$$\begin{aligned}\frac{\partial \kappa_{ij}}{\partial x_{i,k}} &= \frac{\partial}{\partial x_{i,k}}\exp\left(-\frac{1}{2}\left(\mathbf{x}_i-\mathbf{x}_j\right)^{\text{ú}}\mathbf{\Lambda}\left(\mathbf{x}_i-\mathbf{x}_j\right)\right)\\ &= \frac{\partial}{\partial x_{i,k}}\left(-\frac{1}{2}l_k^{-2}\left(x_{i,k}-x_{j,k}\right)^2\right)\kappa_{ij}\\ &= -l_k^{-2}(x_{i,k}-x_{j,k})\kappa_{ij}\end{aligned} \tag{38}$$

where $x_{i,k}$ denotes the k-th feature of $\mathbf{x}_i$.

Based on Equation (37), we can obtain the derivative of the posterior mean of the GP model with respect to $\mathbf{x}_*$,

$$\begin{aligned}\frac{\partial M_*}{\partial \mathbf{x}_*} &= \frac{\partial \boldsymbol{\kappa}_*^{\text{ú}}}{\partial \mathbf{x}_*}\mathbf{A}\\ &= \frac{\partial}{\partial \mathbf{x}_*}\left[\kappa_{1*},\kappa_{2*},\ldots,\kappa_{n*}\right]\mathbf{A}\\ &= -\mathbf{\Lambda}\left[\mathbf{x}_*-\mathbf{x}_1,\mathbf{x}_*-\mathbf{x}_2,\ldots,\mathbf{x}_*-\mathbf{x}_n\right](\boldsymbol{\kappa}_*^{\text{ú}}\circ\mathbf{A})\end{aligned} \tag{39}$$

where $\mathbf{A}=\mathbf{K}^{-1}\mathrm{y}$, and $\circ$ denotes the Hadamard product. Furthermore, the derivative of the posterior mean of the GP model with respect to $x_{*,k}$ is

$$\begin{aligned}\frac{\partial M_*}{\partial x_{*,k}} &= \frac{\partial \boldsymbol{\kappa}_*^{\text{ú}}}{\partial x_{*,k}}\mathbf{A}\\ &= -\mathbf{\Lambda}_k\left[\mathbf{x}_*-\mathbf{x}_1,\mathbf{x}_*-\mathbf{x}_2,\ldots,\mathbf{x}_*-\mathbf{x}_n\right](\boldsymbol{\kappa}_*^{\text{ú}}\circ\mathbf{A})\\ &= -l_k^{-2}\left[x_{*,k}-x_{1,k},x_{*,k}-x_{2,k},\ldots,x_{*,k}-x_{n,k}\right](\boldsymbol{\kappa}_*^{\text{ú}}\circ\mathbf{A})\\ &= -l_k^{-2}\sum_{i=1}^{n}a_i(x_{*,k}-x_{i,k})\kappa_{i*}\end{aligned} \tag{40}$$

where $\mathbf{\Lambda}_k$ is the k-th row of $\mathbf{\Lambda}$, a_i denotes the i-th element of $\mathbf{A}$, and we have $a_i=\sum_{s=1}^{n}\kappa_{is}y_s$.

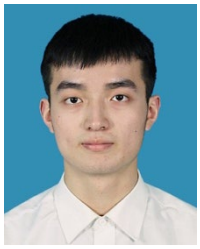
When l_k approaches zero, the GP becomes overfitted in that dimension, indicating a rapid change in the posterior mean.

When l_k approaches infinity, l_k^{-2} will obviously approach zero. $\kappa_{ij,k}$ will also approach zero irrespective of the difference between two input features. Note that the logarithm of κ_{ij} can be written as the sum of m components,

$$\log\kappa_{ij}=-\frac{1}{2}\sum_{s=1}^{m}l_s^{-2}(x_{i,s}-x_{j,s})^2. \tag{41}$$

Hence, the upper bound of κ_{ij} is 1. Accordingly, the value of a_i will not exceed $\sum_{s=1}^{n}y_s$. Thus, we can conclude that $\lim_{l_k\to+\infty}\partial M_*/\partial x_{*_k}=0$. It can be inferred that as l_k becomes larger, the importance of the corresponding feature will decline.

Zhiyuan Liu received his Ph.D. degree in Transportation Engineering in 2011 from National University of Singapore. He is currently a professor at School of Transportation in Southeast University, and director of the Research Center for Complex Transport Networks.
His research interests include transport network modelling, public transport, and intelligent transport system. In these areas, Dr. Liu has published over 70 journal papers.



Cheng Lyu received his B.S. degrees in Transportation Engineering in the School of Transportation at Southeast University, Nanjing, China. He is currently working toward to a M.S. degree in Southeast University.
His research interests include transportation big data analysis and modelling, machine learning, data mining and intelligent transportation systems

Zelin Wang received his B.S. degrees in Transportation Engineering in the School of Transportation at Southeast University, Nanjing, China. He is currently working toward to a M.S. degree in National University of Singapore. His research interests include transportation big data analysis and modelling, intelligent transportation systems and mixed traffic flow of manual-automated driving.

Shuaian (Hans) Wang is a Professor at The Hong Kong Polytechnic University (PolyU). His research interests include shipping operations management, green shipping, big data in shipping, port planning and operations, urban transport network modeling, and logistics and supply chain management. He dedicates to rethinking and proposing innovative solutions to improve the efficiency of maritime and urban transportation systems, to promote environmental friendly and sustainable practices, and to transform business and engineering education.

Pan Liu received the Ph.D. degree in civil engineering from the University of South Florida, Tampa, USA, in 2006. He is currently a Professor with the School of Transportation, Southeast University, Nanjing, China. He has authored or coauthored over 100 articles in transportation journals. His research interests include traffic operations and safety, and intelligent transportation systems. He was a recipient of the Excellent Young Scientist of NSFC in 2019.

Dr. Qiang Meng is currently a Professor in the Department of Civil and Environmental Engineering (CEE) at National University of Singapore (NUS). His research mainly focuses on urban mobility modeling and optimization, shipping and intermodal freight transportation analysis, and quantitative risk assessment of transport operations. He is currently the Co-Editor-in-Chief of Transportation Research Part E and Associate Editor of Transportation Research Part B.